Estimating Gene Regulatory Activity using Mathematical Optimization

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)
im Fach Informatik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät der
Humboldt-Universität zu Berlin

von
M.Sc. Saskia Trescher

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter*innen:
1. Prof. Dr.-Ing. Ulf Leser, Humboldt-Universität zu Berlin
2. Prof. Dr. Katja Nowick, Freie Universität Berlin
3. Prof. Dr. Heike Siebert, Freie Universität Berlin

Tag der mündlichen Prüfung: 30.04.2020

Abstract

Gene regulation is one of the most important cellular processes, indispensable for the adaptability of organisms and closely interlinked with several classes of pathogenesis and their progression, including cancer. The elucidation of regulatory mechanisms can be approached by a multitude of experimental methods, yet integration of the resulting heterogeneous, large, and noisy data sets into comprehensive and tissue or disease-specific cellular models requires rigorous computational methods. Over the last decade, numerous methods have been proposed trying to infer actual regulatory events in a sample. A prominent class of methods models genome-wide gene expression as sets of (linear) equations over the activity and relationships of transcription factors (TFs), genes and other factors and optimizes parameters to fit the measured expression intensities. In various settings, these methods produced promising results in terms of estimating TF activity and identifying key biomarkers for specific phenotypes. However, despite their common root in mathematical optimization, they vastly differ in the types of experimental data being integrated, the background knowledge necessary for their application, the granularity of their regulatory model, the concrete paradigm used for solving the optimization problem and the data sets used for evaluation.

Here, we first review five recent methods of this class in detail and compare them qualitatively with respect to several key properties. Since no comprehensive, comparative evaluation of these methods had been carried out before, we quantitatively compare the results of the presented methods in a unified framework. We base our analyses on different publicly available data sets including TF knockout and knockdown experiments in human and E. coli samples. Our results show that, even in the knockout data sets with clear expression signals and thus an extremely favorable test setting, the mutual result overlaps are very low, though sometimes statistically significant. The knocked out or knocked down TF is rarely identified by any analyzed method. We show that this poor overall performance cannot be attributed to the sample size or to the specific regulatory network provided as background knowledge. However, although drawing very different conclusions when presented with the same inference problem, all methods seem to detect strong signals and, comparing the results to the biological literature, find biologically relevant information. We suggest that a reason for this deficiency might be the simplistic model of cellular processes in the presented methods, where, despite their known importance for gene regulation, TF self-regulation and feedback loops were not represented. We therefore propose a new method for estimating transcriptional activity, named Floræ, with a particular focus on the consideration of feedback loops and evaluate its results in comparison to the previously analyzed methods mainly on synthetic data sets. Using Floræ, we are able to improve the identification of knockout and knockdown TFs in synthetic data sets. Our results and the proposed method extend the knowledge about gene regulatory activity and are a step towards the identification of causes and mechanisms of regulatory (dys)functions, supporting the development of medical biomarkers and therapies.

Zusammenfassung

Die Regulation der Genexpression ist einer der wichtigsten zellulären Prozesse, da sie für die Anpassungsfähigkeit von Organismen unverzichtbar ist und in engem Zusammenhang mit der Entstehung und Entwicklung diverser Krankheiten, unter anderem Krebs, steht. Regulationsmechanismen können mit einer Vielzahl von Methoden experimentell untersucht werden, zugleich erfordert die Integration der daraus resultierenden Datensätze in umfassende gewebe- oder krankheitsspezifische zelluläre Modelle stringente rechnergestützte Methoden. In den letzten zehn Jahren wurden zahlreiche Methoden vorgeschlagen, die die tatsächlichen regulatorischen Ereignisse in einer Probe berechnen. Ein bedeutender Teil dieser Methoden modelliert die genomweite Genexpression als (lineares) Gleichungssystem über die Aktivität und die Beziehungen von Transkriptionsfaktoren (TF), Genen und anderen Faktoren und optimiert die Parameter, sodass die gemessenen Expressionsintensitäten möglichst genau wiedergegeben werden. In verschiedenen Untersuchungen lieferten diese Methoden vielversprechende Ergebnisse zur Identifizierung von zentralen Biomarkern für bestimmte Phänotypen. Trotz ihrer gemeinsamen Wurzeln in der mathematischen Optimierung unterscheiden sich die einzelnen Methoden stark in der Art der integrierten Daten, in dem für ihre Anwendung notwendigen Hintergrundwissen, in der Granularität des Regulationsmodells, im konkreten Paradigma, das zur Lösung des Optimierungsproblems angewendet wird, und in der zur Evaluation verwendeten Datensätze.

In dieser Arbeit betrachten wir zunächst fünf solcher Methoden im Detail und stellen einen qualitativen Vergleich in Bezug auf zentrale Eigenschaften auf. Da bisher keine gemeinsame Auswertung dieser Methoden durchgeführt wurde, führen wir auch einen quantitativen Vergleich der Methoden durch. Unsere Analysen basieren auf verschiedenen öffentlich verfügbaren Datensätzen, unter anderem auf TF-Knockout- und TF-Knockdown-Experimenten. Unsere Ergebnisse zeigen, dass selbst in den Knockout-Datensätzen, in denen deutliche Effekte auf die Expressionsintenstiäten sichtbar sind, die Überschneidungen der Ergebnisse der verschiedenen Methoden untereinander sehr gering, wenn auch in einigen Fällen statistisch signifikant, sind. Der Knockout- oder Knockdown-TF wird nur in den seltensten Fällen erkannt. Wir zeigen, dass diese schlechte Gesamtleistung nicht auf die Stichprobengröße oder das regulatorische Netzwerk zurückgeführt werden kann. Obwohl die Methoden bei gleichen Fragestellungen zu unterschiedlichen Schlussfolgerungen gelangen, scheinen sie gemeinsam dennoch starke Effekte erkennen zu können und finden biologisch relevante Informationen, wie wir im Vergleich mit der biologischen Literatur feststellen konnten. Wir weisen darauf hin, dass die vereinfachten Modelle zellulärer Prozesse ein Grund für die genannten Defizite sein könnten, da diese die vorhandenen Rückkopplungsschleifen, trotz deren bekannter Bedeutung für die Genregulation, ignorieren. Daher schlagen wir eine neue Methode (Floræ) vor, die einen besonderen Schwerpunkt auf die Berücksichtigung von Rückkopplungsschleifen legt, und beurteilen deren Ergebnisse, hauptsächlich anhand synthetischer Datensätze, im Vergleich zu den zuvor analysierten Methoden. Mit Floræ können wir die Identifizierung von Knockout- und Knockdown-TF in synthetischen Datensätzen verbessern. Unsere Ergebnisse und die vorgeschlagene Methode erweitern das Wissen über genregulatorische Aktivitäten und sind ein Schritt in Richtung der Identifizierung von Ursachen und Mechanismen regulatorischer (Dys-)Funktionen, was die Entwicklung von medizinischen Biomarkern und Therapien unterstützt.

Acknowledgments

As much as a dissertation is supposed to be the work of one person, this one might never have gotten finished but for the support and encouragement of some key people. The time as a doctoral student has been a course of intense learning for me, both scientifically and on a personal level. I would like to thank all the people who have supported me during this time.

First and foremost, I want to express my sincere gratitude to Prof. Dr. Ulf Leser for supervising my thesis, for his guidance, help, continuous support and constant motivation throughout the time of my PhD, for the opportunity to visit several conferences, for arranging the extension of my contract, for providing a positive and productive research atmosphere with many ideas, valuable remarks, critical questions and extensive feedback and for being always approachable. My gratitude also goes to Prof. Dr. Katja Nowick and Prof. Dr. Heike Siebert for reviewing my thesis. I thank Prof. Dr. med. Clemens Schmitt and his group members, especially Julia Kase, Kolja Schleich and Animesh Bhattacharya, for the opportunity to work on the Pan-omics project, for proposing interesting lines of research and for giving me a different perspective regarding my thesis. I further acknowledge the funding I received from the DFG via the BSIO graduate school and the CompCancer PhD programme.

I am also grateful to my colleagues of the WBI group at HU Berlin. During the whole time of my PhD, they created a comfortable work atmosphere, provided valuable input to discussions on and off scientific topics, shared their experience and knowledge and asked the right questions. A special thanks goes to Jannes Münchmeyer for being a creative and reliable collaborator.

I thank one of my best friends, Deborah Heinen, for initiating and pursuing our weekly PhD update phone calls. Thank you for your continuous support, for making me stay focused, for reassessing the challenges I faced during my PhD and having great ideas and strategies to overcome them. We were a real winning team and I wish you all the best for the finishing of your thesis.

On a personal note, I thank my parents and my whole family for their constant care, for supporting me in all my undertakings and especially in pursuing my academic career. Most importantly, I owe my deepest gratitude to my husband Max for his genuine and unlimited support, his enormous patience and his trust in me. Thank you for helping and encouraging me whenever I needed it, for providing different and constructive perspectives to my problems, for your love and our wonderful time together as a family with Milo.

Contents

1. Introduction

The discovery of the molecular structure of DNA by [Watson and Crick, 1953] has enabled a fundamentally new approach to the investigation of genetic functions and cellular processes, and in consequence has revolutionized biology and medicine. Not only the structure of DNA and its components have been investigated in more and more detail and in a growing number of species ever since, but also the mechanisms by which genes are expressed, by which gene expression is regulated, and which molecular components are involved are major biological and biomedical research areas [Collins et al., 2003]. The regulation of gene expression is a fundamental biological process, occurring in all living species, which determines the cell's unique properties and enables it to adapt to the organism's development, to cellular function, to the environment and to external stimuli [Spitz and Furlong, 2012]. In eukaryotes, gene expression, i.e. the transcription of genes into mRNA and the subsequent translation into proteins, is mainly regulated by a complex network of transcription factors (TFs), proteins which bind to specific DNA motifs and activate or repress gene transcription. MicroRNAs, which degrade the mRNA transcript, and epigenetic effects, which change the microstructure of the DNA, also influence gene expression. Regulation processes can form feedback loops, and the different mechanisms interact and regulate each other. Alterations or disruptions in regulatory mechanisms can lead to the development and progression of various diseases, including cancer [Beers et al., 2017; Bonder et al., 2017; Naranjo et al., 2016; Semenova et al., 2016]. Thus, the elucidation of regulatory relationships, especially in human, is crucial for the understanding of systematic dysfunctions and the pathogenesis of numerous disorders, constituting an important research field in systems biology [Wang and Huang, 2014].

Over the past years, rapid advances in high-throughput technology and the simultaneous decrease of measurement costs have enabled the investigation of genome-wide expression and other omics data, holding the potential to study biology at systems level [Hogeweg, 2011]. The abundance of large scale data, partly available in public data bases, calls for appropriate analyses methods to enable the understanding of individual cellular entities and their interplay and regulation, represented in gene regulatory networks (GRNs) [Gauthier et al., 2018].

Many algorithms have been proposed to reverse engineer, or infer, the multiple interactions between DNA, RNA, proteins and other cellular molecules directly from expression data [Delgado and Gómez-Vela, 2019]. Compared to individual biological regulator - target experiments, such techniques for structure learning of whole GRNs offer a time and cost efficient way to identify interactions between genes and their products. Network

based approaches are also used to detect the activity of regulators: here, not the formal structure of the GRN itself is searched, but the states of regulatory elements like transcription factors are inferred, resulting in ranked lists of regulators according to their activity. Methods typically integrate prior biological knowledge from experimental evidence of single regulator-target gene interactions, which have been investigated in a very high number of individual biological experiments. Such findings can be retrieved from different databases like TRANSFAC [Wingender et al., 1996] or miRBase [Griffiths-Jones et al., 2006], from integrated resources of different repositories [Garcia-Alonso et al., 2019] or from text mining approaches of publications [Thomas et al., 2015]. Many activity inference methods model genome-wide gene expression as sets of (linear) equations over the activity and relationships of transcription factors, genes and other factors and optimize parameters to fit the measured expression intensities. Some of these methods also integrate different types of omics data to deduce a more comprehensive picture of the key regulatory circuitry acting in a system. Yet none of them considers the effect of feedback loops in the underlying regulatory network, despite their abundance and importance in driving cellular behavior [Brandman and Meyer, 2008; De Jong, 2002; Sauro, 2017]. The authors of the publications presenting activity inference methods claim that such methods can be used to identify biomarkers for specific phenotypes in human cell lines and in vivo samples, for example in innate immunity, ageing related changes [Balwierz et al., 2014] or acute myeloid leukemia [Li et al., 2014]. The type and the extent of evaluation performed for the different methods varies greatly. Although certain evaluation steps were carried out for all methods, the results of the original publications are not comparable as they are based on the evaluation of different data sets using different metrics to assess their performance.

This thesis investigates different methods for estimating gene regulatory activity based on mathematical optimization, quantitatively compares them and proposes a novel method to include the effect of feedback loops. In the following sections, we give an overview of our specific goals and contributions, followed by an outline of the thesis' structure and an account of own prior work in preparation of this thesis.

1.1. Goals and Contributions

The main goal of this thesis is to describe, compare and improve current methods for the estimation of gene regulatory activity based on mathematical optimization. Over the last years, numerous methods trying to infer actual regulatory events in a sample have been proposed, though no comprehensive, comparative evaluation of these methods had been carried out before. Our aim is to reproducibly compare these methods and to identify common shortcomings and necessary extensions, focusing our research on the critical points. To this end, we provide an overview of the state-of-the-art methods in gene regulatory activity estimation, compare their results and performance in different aspects based on several data sets and propose a novel method with a, previously lacking, focus on self-regulation.

The specific contributions of this thesis are as follows:

1. We review five recently published methods for estimating genome-wide gene regulatory activity using mathematical optimization in detail, namely the approach by [Schacht et al., 2014], RACER [Li et al., 2014], RABIT [Jiang et al., 2015], ISMARA [Balwierz et al., 2014], and biRte [Fröhlich, 2015]. We compare these methods qualitatively with respect to several key properties with the goal to identify their mutual strengths and weaknesses. All methods produce a ranked list of TFs, sorted by their activity in a given group of samples. They differ in the types of measurements being integrated, the background knowledge necessary for their application, the complexity and refinement of the underlying model of gene regulation and the concrete paradigm used for solving the optimization problem. We emphasize the common ground of these at-first-sight rather different methods by explaining similarities and differences to a general framework for defining the relationships of transcription factors and genes.

2. Although evaluations were carried out for all methods in the original publications, the results are not comparable as they used different input data sets, different background regulatory networks and different evaluation metrics. We implement a quantitative comparison to objectively analyze the results of the previously presented methods for estimating regulatory activity in a unified framework and further to investigate the influence of the network topology on the results. We base our analyses on publicly available data sets including different regulator - gene networks, multi-omics experimental patient data and transcription factor knockdown experiments from human and E. coli cell lines to ensure transparency and reproducibility of our results.

3. We propose Floræ (Feedback loops in regulatory activity estimation) as a novel method for estimating regulatory activity with a particular focus on the consideration feedback loops in the underlying gene regulatory network. Floræ is constructed modularly to facilitate the adaptation to different applications and contexts. To allow the control of all parameters, we evaluate the results in comparison to the previously analyzed methods using mainly synthetic data, simulating knockout and knockdown experiments. We further examine the influence of the network's topology and the number of samples on the results and apply Floræ to real biological data.

1.2. Outline

Here, we give a brief overview of the structure and content of this thesis in a chapter-based manner.

Chapter 2 introduces basic (biological) concepts relevant throughout this thesis. It provides an overview of gene expression and gene regulation mechanisms with a focus

on transcription factors and microRNAs. We present models of gene regulatory networks and describe methods for their reconstruction as well as current challenges in this research field. Further, Chapter 2 introduces the idea of gene regulatory activity inference, a variation of network reconstruction, and mathematical optimization as popular method to infer regulatory activity.

Chapter 3 surveys and qualitatively compares different methods for estimating genome-wide gene regulatory activity. We introduce a general mathematical framework for the inference of regulatory activity and describe five published methods in detail, focusing on incorporated data types, mathematical models, optimization methods, and evaluation strategies. We descriptively compare the general properties of these methods and discuss their strengths and weaknesses, motivating the necessity for detailed quantitative comparisons based on controlled data scenarios.

Chapter 4 studies our results on the quantitative evaluation of different methods for estimating regulatory activity. We report on method configurations and the ranking procedure. The extensive comparison is divided into two sections and is based on different publically available data sets: multi-omics data from cancer patients and knockdown data from human and E.coli cell lines. For the analyses based on multi-omics data, we comparatively report on our results using different amounts of input data sets. In a second evaluation based on less complex knockdown experiments, we compared the results from different methods and analyzed the influence of the underlying regulatory network. We discuss our results with respect to study design and network topology, pointing out the need for the inclusion of self-regulation in activity inference methods.

Chapter 5 proposes a novel method for the estimation of regulatory activity, Floræ, with a particular focus on the consideration of feedback loops in the regulatory network. We first describe the methodological background and the implementation of Floræ. We report on our evaluation based on synthetic data and the data sets derived from biological experiments described in Chapter 4. The effects of feedback loops, sample number and network randomization are examined in detail. We discuss the advantages and limitations of the methodological approach and of the use of synthetic data, indicating several potential extensions of Floræ.

Chapter 6 summarizes the results of this thesis, recapitulates the main contributions and addresses possible future directions.

1.3. Own prior Work

Some parts of this thesis are based on work which has been published previously in peer-reviewed publications. Chapters 3 and 4 describe the comparative assessment of methods for estimating regulatory activity based on multi-omics data originally presented in [Trescher et al., 2017]. Saskia Trescher performed the literature research, the

quantitative comparisons and wrote the manuscript with the help of Jannes Münchmeyer and Ulf Leser. Chapter 4 further contains the analyses of transcription factor activity in knockdown studies which were published in [Trescher and Leser, 2019]. Saskia Trescher performed the literature research, implemented the in silico experiments and analyzed the data. Saskia Trescher wrote the manuscript together with Ulf Leser.

2. Background

The regulation of gene expression is a fundamental biological mechanism in all living cells. It determines the cells' unique properties and it is indispensable for the organism's development, cellular function and the adaptation to changing environments and external stimuli [Spitz and Furlong, 2012]. Gene regulation also plays an important role in the development and progression of various diseases [Jargosch et al., 2016; Kleinjan and van Heyningen, 2005; Maurano et al., 2012]. Further, the distortion of regulatory processes is inflicted with various diseases [Gong et al., 2010; Jiang et al., 2009], especially with cancer [Esquela-Kerscher and Slack, 2006; Mayo and Baldwin, 2000].

The following chapter presents the essential biological and technical background for the understanding of the remainder of this thesis. First, we describe the biological mechanisms of gene expression, including the technical bases of determining gene expression by microarray and state of the art high-throughput sequencing technology. Subsequently, the role of transcription factors, post-transcriptional regulation and epigenetics in gene regulation is described. We introduce the concept of gene regulatory networks and describe different methods for their reconstruction. Furthermore, we present basic notions of the inference of regulatory activity and thus the elucidation of regulatory relationships based on mathematical optimization, as studied in this thesis.

2.1. Gene Expression

Gene expression refers to the process of transcribing a specific segment of deoxyribonucleic acid (DNA), called gene, to ribonucleic acid (RNA) and the subsequent translation into a functional gene product (see Figure 2.1). The procedure consists mainly of three steps: First, DNA is transcribed into RNA followed by a splicing step, where the non-coding regions (introns) are removed from the RNA whereas the coding regions (exons) are joined together, forming the messenger RNA (mRNA). Finally, the mRNA is translated into an amino acid sequence, which folds into a functional protein [Alberts et al., 2014]. Further, a large number of non-coding RNA exist, which are not translated into proteins and which are involved in many cellular processes, having different, partly unknown functions [Washietl et al., 2007].

In prokaryotic organisms, which lack a defined nucleus, the DNA floats freely within the cytoplasm, and the processes of transcription and translation occur almost simultaneously. In contrast, in eukaryotic cells the processes of transcription and splicing (in the nucleus) and translation (in the cytoplasm) are physically separated by the nuclear membrane [Kozak, 2005]. In this thesis, we will analyze gene expression measurements,

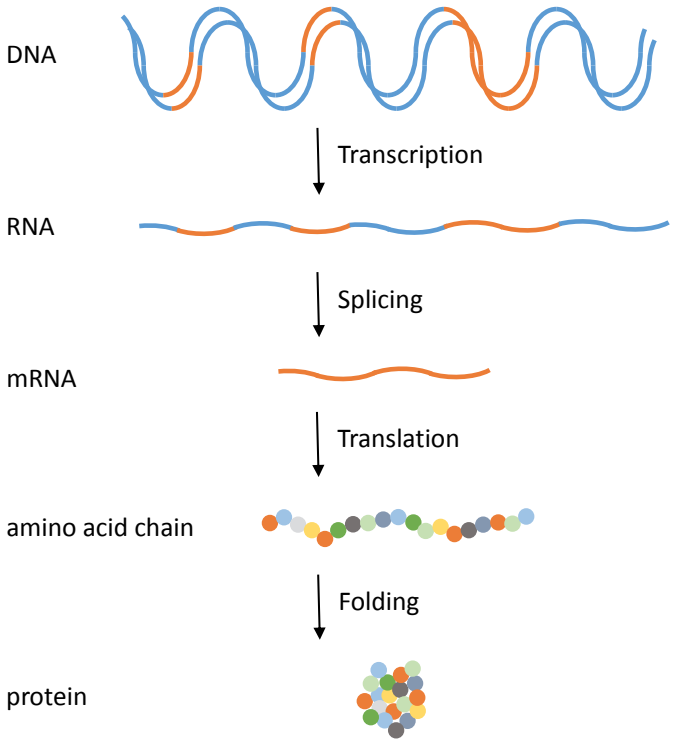


Figure 2.1.: Process of gene expression including transcription of DNA into RNA, splicing into mRNA, translation into an amino acid chain and folding into a protein.

both of prokaryotic organisms like the bacteria E.coli and eukaryotic cells from human cell lines and patient samples.

Experimental Technologies

For most bioinformatic analyses, gene expression is measured at high-throughput scale, where the mRNA levels of many genes within a sample are analyzed simultaneously. The two main technologies for sensing gene expression are hybridization microarrays [Schena et al., 1995] and next-generation sequencing (NGS) such as RNA-Sequencing [Voelkerding et al., 2009]. Methods of comparative gene expression analysis or gene expression tracking over time using microarray or sequencing technology may help to elucidate regulatory mechanisms at transcriptome level. In Chapter 3, we will describe methods analyzing gene expression data originating from microarrays or RNA-Seq and later use such measurements for our own computational analyses in Chapters 4 and 5.

Microarrays are based on the principle of hybridization between two complementary DNA strands. Defined genetic sequences (probes) are attached to specific locations of the two dimensional surface of a chip. These probes consist of short nucleotide sequences

ideally representing unique sequences of the gene they target. By adding the sample's transcriptome to the chip, the mRNA complementary to the gene-representing probes hybridizes to the latter and can then be quantified by the activation of an fluorescent labeling and subsequent optical recognition of the amount of bound sequences for each spot.

The standard protocol of performing a microarray expression experiment consists of several steps [Cheung et al., 1999] (see Figure 2.2). First, the target cell's mRNA is extracted and purified to avoid false positive signals through contamination. Via reverse transcription, complementary DNA (cDNA) is generated from the mRNA. These cDNA fragments are then labeled, typically with fluorescent dye, and added to the microarray chip, where they bind to the complementary sequences of the probes. After this hybridization step, residual unbound sequences and DNA fragments are washed off. By scanning the emitted light of the previously coupled fluorescent dye, the amount of hybridized mRNA is captured. Numerical quantities from the dye intensities in the picture are computed and normalized to reduce the technical bias between and within arrays [Brazma et al., 2001]. The normalized expression values can then be analyzed, for example in statistical tests for expression differences between sample groups, clustering algorithms, classification methods or for gene network inference.

Microarrays represent a well established and low-cost technology for measuring gene expression [Lee et al., 2008], costing about 100$ per sample [Yandell, 2015], also to the present time. However, they cannot quantify the exact amount of mRNA in the cell, but only the amount of abundant transcripts depending on binding specificities or saturation. Further, microarrays are limited to measuring known genes, as their complementary sequence has to be attached to the array in advance [Zhao et al., 2014].

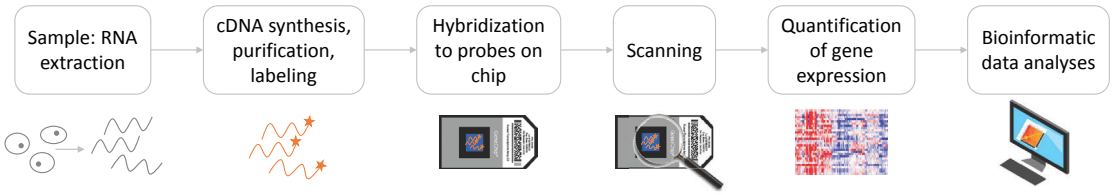


Figure 2.2.: Schema illustrating the processing steps and their sequential order of a microarray experiment from sample RNA extraction to data analysis.

Another method for measuring mRNA expression is high-throughput RNA-Sequencing (RNA-Seq). The number of detected sequences, called reads, matching a specific coding region in the genome allows to quantify mRNA expression levels [Reuter et al., 2015]. A widely adopted NGS technology is the sequencing-by-synthesis approach detecting single bases as they are incorporated into growing DNA strands [Fuller et al., 2009] (see Figure 2.3). First, the mRNA of interest is extracted from the cells, fragmented into smaller pieces and reverse transcribed to cDNA. The resulting double-stranded cDNA fragments

are ligated with adapter sequences on both ends. This so-called library is then attached to the surface of a flow cell, a glass slide with eight flow channels. Each bound fragment is amplified by synthetization of the complementary strand. This leads to the building of clonal clusters, as many copies of one fragment are located in close spatial proximity. Sequencing reagents including the four fluorescently labeled nucleotides are added and incorporated in the sequence. The flow cell is imaged and the emission wavelength and intensity are used to identify the incorporated base. By repeating the sequencing cycles, the sequence structure can be reconstructed. The reads are computationally aligned to a reference sequence and can be used for further analyses, like expression quantification or the detection of differentially expressed entities.

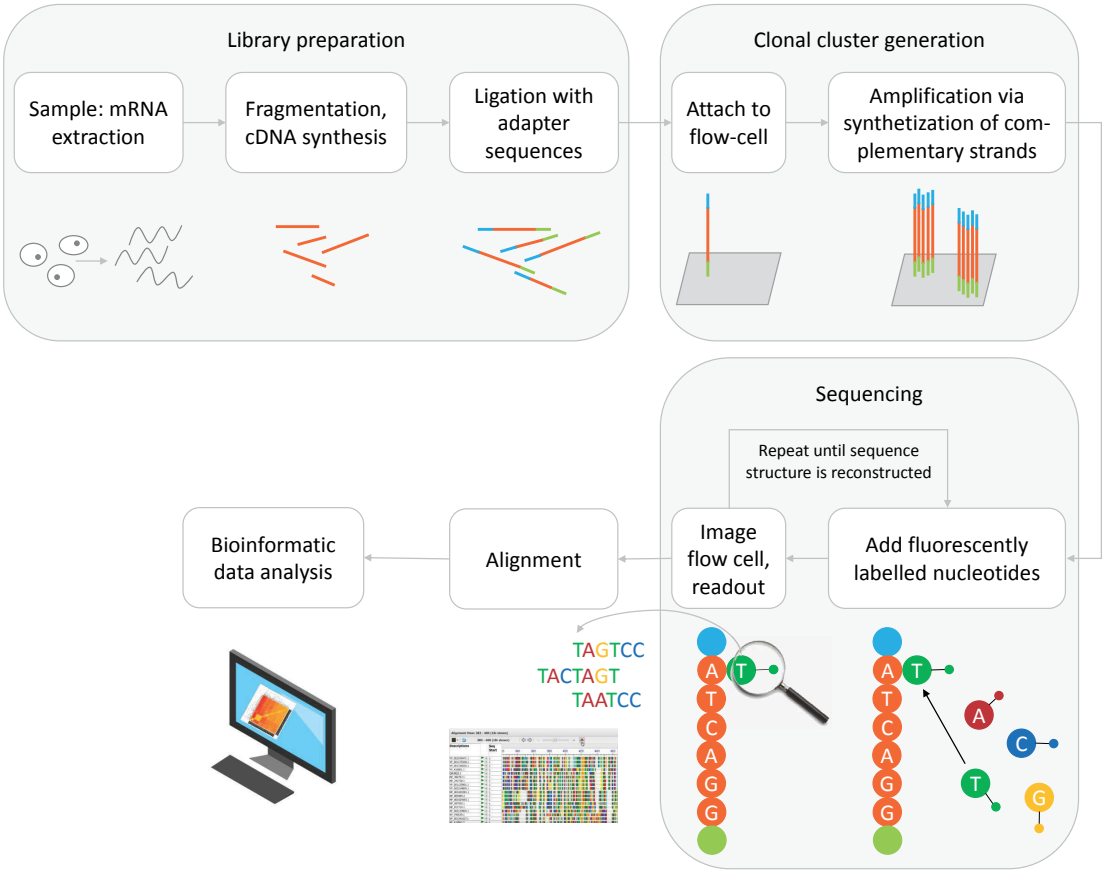


Figure 2.3.: Schema illustrating the processing steps and their sequential order of a RNA-seq experiment from sample RNA extraction to data analysis.

Unlike microarrays, RNA-Seq technology is able to capture sequences not previously known, as it is not dependent on pre-designed probes present on the chip [Zhao et al., 2014]. While for microarrays expression measurement is limited by signal saturation at the high and background noise at the low end, RNA sequencing provides discrete

read counts and thus a better qualitative and quantitative transcriptome acquisition [Nagalakshmi et al., 2008]. For a long time, the main disadvantage of RNA-Seq was the high cost, but over the years, sequencing technology and necessary computation power became cheaper by magnitudes. Currently, sequencing a human genome with a size of around 3,000 Mb costs slightly above 1,000$ [Wetterstrand, 2019]. While for instance Illumina indicates its average error rate of sequence reads with 1% per base, other solutions exist, which show a much higher error rate of 5%-40% with the advantage of a significant speedup [Goodwin et al., 2015]. Further, RNA-Seq sample preparation and computational analysis routines are not yet standardized [Nekrutenko and Taylor, 2012], making it difficult to compare technologies and results.

2.2. Gene Regulation

The regulation of gene expression is essential to the functions and mechanisms of organisms, like their development, reaction to external stimuli or the adaption to the environment [Alberts et al., 2014]. Gene regulation refers to all mechanisms cells use to increase or decrease the creation of specific gene products. In eukaryotes, modulation happens at every step of gene expression, for example through structural or chemical DNA modification, transcriptional control during transcription of DNA to RNA, at the control of RNA processing, during the transport from the nucleus to the cytoplasm, via mRNA degradation, or during the translation into a protein. Regulatory effects result from a complex interplay of multiple of these mechanisms. Further, it is assumed that more currently unknown factors are involved in gene regulation and that not all regulatory effects can be fully explained by now [Munsky et al., 2012]. In this thesis, we mainly focus on the regulatory effects of transcription factors, but also microRNAs and epigenetic effects like DNA methylation are considered in some of the presented methods for estimating regulatory activity.

2.2.1. Transcription Factors

The transcription of DNA into RNA is predominantly controlled by a complex network of transcription factors (TFs) [Alberts et al., 2014]. These proteins bind to distal or proximal binding sites at characteristic sequence motifs of DNA adjacent to the genes they regulate [Lemon and Tjian, 2000], which may enhance or inhibit the recruitment of RNA polymerase and thereby activate or repress gene transcription [Spitz and Furlong, 2012] (see Figure 2.4). Approximately 10% of all human genes code for TFs, making them the largest family of human proteins [Jolma et al., 2013]. In this work, we mainly focus on TFs as origin of regulatory events, as they are assumed to be one of the most important factors during gene regulation [Alberts et al., 2014].

The estimated number of TFs in the mammalian genome is about 1500-2600, but only half of them is known [Babu et al., 2004; Vaquerizas et al., 2009]. Several human diseases have been associated with TF mutations [Lambert et al., 2018b]. An aberrant regulation of TFs, that act as oncogenes or tumor suppressors, is associated with cancer. For

example, the STAT TF family is involved in the oncogenesis of breast cancer [Clevenger, 2004] and HOX TFs are associated with kidney and colon cancer [Cillo et al., 1999]. Therefore, TFs are of high clinical significance since they can be direct targets of medications or indirectly regulated through signaling cascades [Bhagwat and Vakoc, 2015; Gronemeyer et al., 2004; Lambert et al., 2018a]. For example, the MYC TF family is deregulated in more than 50% of human cancers [Chen et al., 2018]. MYC transcription and translation inhibition have been studied to target this TF for cancer therapeutic purposes, e.g. via the inhibition of CDK7, an essential component of the transcription factor TFIIH [Chipumuro et al., 2014].

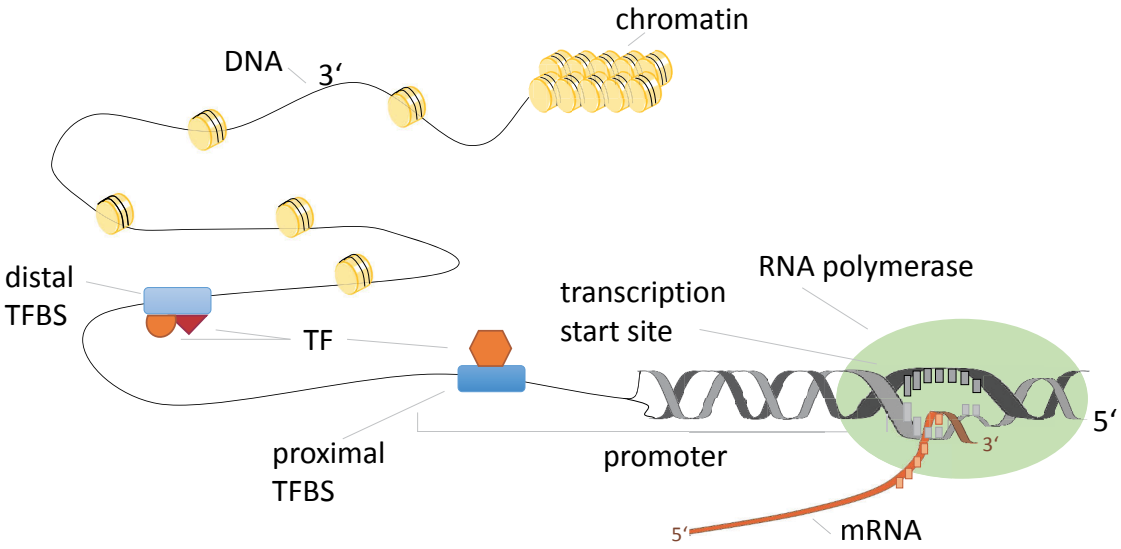


Figure 2.4.: Gene regulation via transcription factors. Transcription factors (TFs) bind to distal or proximal TF binding sites (TFBS) enhancing the binding of RNA polymerase and activating the transcription of DNA into RNA.

In recent years, chromatin immunoprecipitation followed by sequencing (ChIP-seq) has become the gold-standard for the detection of TF binding sites (TFBS) and for profiling the binding of transcription factors to DNA at a genome-wide scale [Furey, 2012]. Such experiments provide hundreds to thousands of potential binding sites for a given transcription factor in proximity to gene coding regions [Lachmann et al., 2010]. Technically, chromatin is chemically fixated with formaldehyde and the DNA and TF of interest are co-precipitated using an antibody targeting that TF. The bound DNA sequences can then be identified by high-throughput sequencing. Further, computational methods are used to predict new TFBS [Johnson et al., 2007] and to find known TFBS within the genome (e.g., [Elemento and Tavazoie, 2005; Ernst et al., 2010]). Several databases have been created which store relevant information, such as lists of binding motifs (TRANSFAC [Wingender et al., 1996] or JASPAR [Sandelin et al., 2004]). However, since each

ChIP-seq experiment is limited to the detection of one TF in one condition and since TFs may be active simultaneously, TF binding has not been characterized yet comprehensively for many TFs in different cell types. The in Chapter 3 described methods for estimating regulatory activity incorporate knowledge about TF binding in different ways. We later will use knowledge about TF – gene interactions from TRANSFAC for our analyses of TF activity (see Chapters 4 and 5).

TFs can play antagonistic roles in the regulation of the same gene, when they compete for binding to a specific TFBS [Teif and Rippe, 2010]. TFs can also bind in a combinatorial manner, allowing genes to be regulated in complex patterns in both space and time [Spitz and Furlong, 2012]. Further, TFs not only control the regulation of genes, but also the production rate of other transcription factors or even themselves, called autoregulation. For example, a TF can form positive or negative feedback loops, acting as inducer or repressor for other TFs including itself [Sankpal et al., 2017]. Self-regulation can enable the cell to maintain a high or low level of a certain TF [Pan et al., 2006]. Since feedback is an important mechanism in gene regulation, we later will introduce a method for estimating transcriptional activity with a particular focus on the consideration of feedback loops (see Chapter 5).

2.2.2. MicroRNAs

For post-transcriptional regulation, microRNAs (miRNAs) play a major role. These small non-coding RNA molecules function via base-pairing with complementary mRNA sequences. They act on gene regulation directly by degrading the mRNA transcript or indirectly by inhibiting their translation [Guo et al., 2010]. Most miRNAs alter the protein expression of their target genes only modestly by a factor of 1.5 to 4 [Farazi et al., 2011]. In the human genome, around 1200 different miRNAs are known, playing a role in the regulation of more than 30% of all known mRNAs [Rajewsky, 2006] , e.g. in circadian regulation [Lehmann et al., 2015]. In this thesis, miRNAs are considered, next to TFs, in some of the later presented methods as important regulators whose activity can be estimated (see Chapter 3). We will use miRNA-gene networks, indicating which miRNAs are able to degrade which mRNA transcript, to analyze the activity of miRNAs in cancer (see Chapter 4).

As well as TFs, miRNAs can act as tumor suppressors in cancer [Esquela-Kerscher and Slack, 2006]; generally their de-regulation has shown to play a role in various diseases [Jiang et al., 2009]. For example, as a deficiency of BRCA1 can cause breast cancer [Magdinier et al., 1998], increased expression of miR-182 down-regulates BRCA1 expression, and increased miR-182 is found in 80% of breast cancers [Krishnan et al., 2013]. Alterations in microRNAs often down-regulate DNA repair mechanisms, which represents an important step in cancer pathogenesis and progression [Hatano et al., 2015]. MiRNA-based therapies are currently investigated in clinical trials [Ganju et al., 2017; Romano and Kwong, 2018; Takahashi et al., 2019].

MiRNA expression can be experimentally detected via hybridization to miRNA microarrays or by high-throughput sequencing. However, high-throughput quantification of miRNAs is error prone, since miRNAs degrade more easily due to their short length, and have a higher variance compared to mRNAs, leading to sample preparation and methodological problems. Relevant information on targets of regulatory miRNAs is, for example, stored in miRBase [Griffiths-Jones et al., 2006]. Computational approaches paring mRNA and miRNA by predicting miRNA-targets based on their sequences exist as well, however, it has been suggested that many functional miRNAs are missed by target prediction algorithms [Nourse et al., 2018].

2.2.3. Epigenetics

Epigenetics refers to changes of the microstructure of the DNA or the associated chromatin proteins, which are heritable and functional, but do not entail changes in the DNA sequence itself [Wu and Morris, 2001]. Epigenetic effects cause activation or silencing of certain genes via two major mechanisms: Histone modifications and DNA methylation [Jaenisch and Bird, 2003]. Histone modifications change the shape of the histones and thus the DNA wrapping around them, possibly leading to gene expression changes. For example, histone acetylation converts the positively charged amine group of the histone into a neutral amide linkage. This removes the positive charge, thus loosening the DNA from the histone. Subsequently, TFs can bind to the DNA and allow transcription to occur [Clapier and Cairns, 2009; Khan, 2014]. However, the predominant epigenetic modification in mammalian DNA is methylation of cytosine in CpG dinucleotides [Kim et al., 2008]. Highly methylated areas of DNA tend to be less transcriptionally active by preventing the binding of transcription factors [Watt and Molloy, 1988]. In Chapters 3 and 4, we will describe and evaluate the use of DNA methylation data as input for methods estimating regulatory activity.

Epigenetic changes contribute to the genesis of different diseases like cancer, coronary heart disease, stroke, diabetes or developmental diseases [Dupont et al., 2009; Mamlouk et al., 2017]. For example, hypermethylation at the promoter CpG islands of a tumor suppressor gene allows cells to grow and reproduce in an uncontrolled manner, leading to tumorigenesis [Esteller, 2007; Gopalakrishnan et al., 2008]. More than 90% of prostate cancers show gene silencing by CpG island hypermethylation of the GSTP1 gene promoter, which normally protects prostate cells from genomic damage [Gurel et al., 2008]. Further, cancer cells tend to have less monoacetylated and trimethylated forms of histone H4 compared to healthy cells [Fraga et al., 2005]. Therefore, manipulating epigenetic changes is highly interesting for cancer prevention and therapy. For example, the DNA methyltransferase inhibitors azacitidine [Garcia-Manero, 2008] and decitabine [Aribi et al., 2007] target the distorted methylation pattern of cancer cells and are used in the treatment of a specific blood cancer, myelodysplastic syndrome.

To find epigenetic signals of regulation, high-throughput genome-wide analysis of DNA methylation is conducted via bisulfite sequencing, which is considered the gold standard

for measuring CpG methylation [Lou et al., 2014]. During bisulfite conversion, unmethylated cytosine is converted into uracil. Subsequent sequencing and re-alignment to the reference genome allows the detection of mismatches and therefore the methylation states of CpG dinucleotides [Chatterjee et al., 2012]. Information on DNA methylation and its patterns in different species and conditions are for example stored in Pubmeth [Ongenaert et al., 2008], iMETHYL [Komaki et al., 2018] or MethBank 3.0 [Li et al., 2018].

2.2.4. Further mechanisms

TF binding itself is affected by chromatin state [Kasowski et al., 2013]. DNA packed in nucleosomes is generally inaccessible to transcription factors, and only unwrapped DNA allows access to the transcription factor binding site. It is also unlikely, that a TF binds to all matching DNA sequences identified by ChIP-Seq in vivo, since DNA accessibility or the presence of co-factors might influence the actual binding, making it a difficult task to predict where a TF will actually bind in a living cell.

Further, TFs can bind to a distal promoter (enhancer or silencer), up to one megabase pairs distant from the gene they regulate. In such cases, the DNA strand bends such that the TF is spatially close and is able to bind to the core promoter. This effect can be detected via chromosome conformation capture (3C) technologies, quantifying the number of interactions between genomic loci [de Wit and de Laat, 2012]. Via high-throughput sequencing of the interacting loci (Hi-C), it is possible to analyze genome-wide chromatin organization [van Berkum et al., 2010]. These methods have revealed a large-scale organization of the genome into topologically associating domains (TADs), in which DNA sequences interact more frequently with each other compared to sequences outside the TAD [Pombo and Dillon, 2015; Rao et al., 2014].

In this thesis, we will not use data of histone modifications describing DNA accessibility and ignore co-factors of transcriptional regulation and the influence of distal promoters. However, these mechanisms present interesting possibilities for further research on the estimation of regulatory activity (see Chapter 6).

2.3. Gene Regulatory Networks

During the last decades, the biological knowledge about genes, proteins and their interactions was constantly growing and required the development of adequate forms of representation and analysis. Gene regulatory relationships are often represented in networks, as they enable an intuitive characterization of complex biological mechanisms. Today, the elucidation of regulatory relationships on large scales is one of the most important goals in systems biology. In this thesis, we consider gene regulatory networks as a key component for the inference of regulatory activity.

2.3.1. Model

Gene regulation, as described in Section 2.2, consists of complex interactions between molecular entities, leading to e.g. gene activation or self-regulation. To describe biological reality and predict the behavior of biological signaling systems, it is necessary to build abstract models. Typically, gene regulation is modeled in pathways, which describe a linked series of interactions of molecules in a cell leading to downstream responses and which can affect each other [Jin et al., 2014]. Starting from a detailed description of chemical processes and intermediate signaling molecules, pathways can be abstracted further in different granularities, concentrating on the entities of interest, like TFs, genes and their interactions [De Jong, 2002]. These TF - gene interactions can represent a direct physical binding of a TF to a target gene, but might also comprise an indirect relationship, for example when the expression of a directly regulated gene in turn influences the expression of others, or when a regulation is caused by one or more intermediaries. Integrating several pathways into a single formalized model, leads to their description as a graph, which is called gene regulatory network (GRN) [Bolouri and Davidson, 2002]. Such a graph can be visualized as a network to depict the coherence of biological entities.

A graph G is defined by the pair $(V(G), E(G))$ where $V(G)$ denotes the set of nodes and $E(G) \subseteq V(G) \times V(G)$ denotes the set of edges. An edge $e \in E$ is defined by two adjacent nodes $i, j \in V$. In a directed graph, an edge is defined by an ordered pair of nodes (i, j) denoting the edge direction, pointing from node i to node j. Weighted graphs are defined as $G = (E, V, w)$ where $w : E \to W$ is a mapping function of edges to discrete or continuous values (weights). A path in a graph is a sequence of nodes joined by a sequence of edges.

In a GRN, genes, TFs, miRNAs and other cell components can be conceptualized as nodes, and their interactions as edges [Steele et al., 2009]. The nodes can be categorized into regulators (e.g. TFs) and regulated entities (e.g. genes). Directed edges indicate a regulatory relationship between the two connected nodes, for example the influence of a TF on the expression of a gene. Edges in GRNs can be weighted to attribute a quantitative measure to a regulatory interaction. In a simple case, the edge weights are $w \in \{-1, 1\}$ to indicate that a TF is inhibiting or activating the expression of a target gene [Hecker et al., 2009]. A given regulator may have hundreds of different targets, and a given target might be regulated by multiple regulators (see Figure 2.5). In chapter 4, we will use for example a human TF - gene network and a miRNA - gene network to infer regulatory activity in cancer. As regulators might not only influence the regulation of genes, but also the production rate of other regulators or themselves, loops are frequent structures in GRNs [Milo et al., 2002]. We represent loops in the TF - gene network as edges from a TF to a gene and vice versa, from a gene to a TF (compare Figure 2.5), like in [Kel et al., 2019], [Isomura and Kageyama, 2014] or [Vlaic et al., 2012], meaning that a TF binds to a target gene to influence its expression, and that the target gene produces a protein or signaling molecule, that plays a role in the cascades regulating the activity of the TF. We developed a method for estimating TF activity with a focus on feedback loops in the network (see Chapter 5).

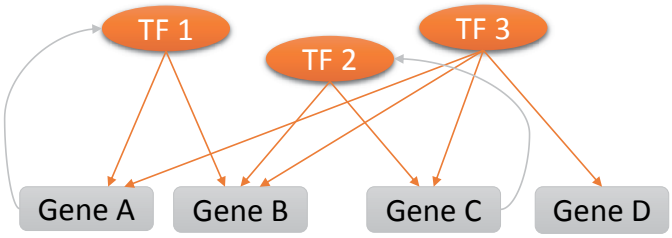


Figure 2.5.: General scheme of a gene regulatory network including TFs and genes. Orange edges indicate a regulatory relationship between a TF and a gene via TF binding. Gray edges indicate the production of proteins by the gene, which act on the formation or decomposition of TFs, forming feedback loops.

GRNs can be extended to model various levels of biological data from gene regulation and protein interaction to metabolic and biochemical reactions. In this thesis, we mainly focus on the interactions of TFs and genes, and do not include the temporal evolution of molecular interactions and epigenetic effects in the pathway model here. Further, co-factorial binding and compound-building of regulators are not considered. To include such concepts, the use of hypergraphs, that are able to capture many-to-many relationships, would be necessary [Bolouri, 2014].

2.3.2. Reconstruction

GRN reconstruction aims at the identification of regulatory mechanisms. In this work, the term network reconstruction (or inference) describes the process of computationally predicting direct and indirect interactions between regulatory elements, such as activation, inhibition or binding, based on biological experimental evidence.

GRN reconstruction is a powerful tool for deciphering regulatory interactions, but its performance is highly dependent on the quality and amount of available input data. Once a network is reconstructed, it is difficult to evaluate its quality, since there are no gold standard networks for higher organisms like mammals. Many evaluations of reconstruction algorithms are based on artificial data, which are not adding to the understanding of biological networks [Thomas and Jin, 2014]. Further, reconstruction algorithms themselves suffer from the dimensionality "large p, small n" problem, referring to the high number of regulatory effects, that should be modeled, compared to a small number of samples providing biological data. We will discuss these limiting issues in greater detail in Section 2.3.3.

GRN reconstruction algorithms use different techniques to circumvent these problems, and many approaches of GRN reconstruction have been published over the years. Computational tools operating on gene expression or other high-throughput data allow the reconstruction of GRN networks in a time and cost efficient manner [Wang and Huang,

2014]. These methods have been successfully used, for example, in the diagnosis of hepatocellular carcinoma [Liang et al., 2018] or the identification of biomarkers in cancer progression and treatment [Yan et al., 2016]. The methods can be categorized into those who reconstruct static gene networks using steady-state data, and those who infer dynamic networks based on time-series data to reflect temporal changes of gene regulation. Here, we give an overview of methods for reconstructing the structure of a GRN and later explain the possibility of estimating regulatory activity in Section 2.3.4.

Static networks

In static network reconstruction, experimental data is measured at a fixed point in time to infer regulatory activity, for example, in drug-response scenarios or different cell types.

A popular method is the construction of **co-expression networks**: Gene-gene relationships are predicted whenever the correlation between both genes in a sample is above a certain threshold. It is assumed, that genes with similar expression profiles under different experimental conditions are likely to be co-regulated and hence functionally related [Wang and Huang, 2014]. Correlation can be determined via coefficients like the Pearson or Spearman correlation, or the Euclidean distance [D'Haeseleer et al., 2000]. It is further possible to assign the correlation value to each edge in the network and thereby obtain a weighted network [Butte and Kohane, 2013]. As an example, co-expression networks were successfully applied in the discovery of conserved genetic modules across evolution [Stuart et al., 2003]. Co-expression networks are easy to interpret and to construct computationally, even in the case of low gene expression or a small number of samples. However, the regulation by multiple genes is not considered and it is a major challenge to define an adequate correlation threshold and to choose a suitable correlation measure. Further, co-expression does not necessarily indicate a regulatory relation, thus leading to high false positive and low prediction rates [Gillis and Pavlidis, 2012].

Compared to correlation coefficients, a more general way to measure gene relationships is the information theoretic measure **mutual information** (MI). A MI equal to zero indicates that the ensemble of two genes do not contain more information than both on their own. After a discretization step, the entropy for each variable and pairwise MI is calculated, which in turn is used to infer the GRN. MI is, unlike correlation coefficient measures, shown to be able to capture non-linear correlations between expression profiles [Daub et al., 2004]. For example, ARACNE [Margolin et al., 2006]and CLR [Faith et al., 2007] are popular GRN reconstruction methods based on mutual information (see Chapter 3).

Co-expression and MI based reconstruction algorithms only consider pairwise relations between the nodes of the network and might miss higher-level interactions. A gene might simultaneously interact with a group of genes, without having a dominant relationship with any individual gene in the group. **Gaussian graphical models** (GGM) can be used to represent conditional dependencies between nodes and allow to distinguish be-

tween direct and indirect associations. Based on gene expression data, GGMs were, for example, used to identify disease candidate genes of multiple sclerosis [Li et al., 2007a]. Despite their favorable theoretical properties, the quality of the inferred network depends highly on the correct selection of a set of genes, on which the correlation is conditioned [Kim et al., 2012]. Different heuristics to find an optimal set were proposed [Chu et al., 2009; Glymour et al., 2019; Opgen-Rhein and Strimmer, 2007].

The aforementioned methods construct undirected graphs and cannot represent underlying biological causal relationships between genes. **Bayesian networks** [Friedman et al., 2000] are popular GRN inference methods for constructing directed acyclic graphs. Each node in the network is treated as a random variable, whereas the graph represents the joint probability distribution of all nodes [Chai et al., 2014]. The reconstruction encompasses two steps: Learning the graph structure given the observed gene expression data, and subsequently learning the parameters of local conditional probabilities given the graph's structure. To reduce the search space for the structure of the graph, it is possible to include biological assumptions and priors. To identify the Bayesian network that best fits the given data among all possible ones, a scoring function is considered [Heckerman et al., 1995; Konishi et al., 2004]. Bayesian models were for example applied in the identification of miRNAs in kidney cancer [Chekouo et al., 2015]. Bayesian networks provide a flexible setup for gene network inference, as they allow the inclusion of prior knowledge or latent variables. However, this increases the number of parameters and therefore the need of a growing amount of high quality data. Further, feedback loops cannot be included and the network size is limited by computation power, as the associated search space of structure learning is superexponentially large [Lucas, 2004].

Dynamic networks

To capture the dynamic behavior of real networks, such as different states during the temporal course of a biological process, it is necessary to collect measurements over time and apply dynamic network inference methods like Boolean networks, dynamic Bayesian networks or models based on differential equations.

Boolean networks, which were proposed quite early [Kauffman, 1969], are suitable to model interactions and causal relations between nodes, for example to describe oscillations or switch-like behavior stability [Delgado and Gómez-Vela, 2019]. For each node, the expression level is discretized to two states and its changes of state between different time points is described by a Boolean function of its parents nodes. Reconstruction of the network is achieved by composing directed graphs where the nodes are connected to each other by means of Boolean functions. This set of functions in effect determines a topology on the set of nodes, constituting the regulatory network. Boolean networks are easy to implement and work quite well even without prior knowledge or with small amounts of input data, for example, for simulating GRNs [Chai et al., 2014]. [Moignard

et al., 2015] used a Boolean network to reconstruct a model for blood development, which was later experimentally validated. The main limitation of Boolean networks lies in the discretization step, making them unable to capture quantities or complex behaviors of real world systems like decreases and increases of gene expression. Further, the time span between different time points can not be modeled, and different time points describe only a sequence of consecutive states.

Bayesian networks can be extended to capture temporal relationships between variables and thus modeling loops [Friedman et al., 2000]. Directed acyclic graphs are generated for each time point as described for Bayesian networks while parents of a node can include nodes from previous time points. By merging the graphs, **dynamic Bayesian networks** can model circles or loops which originally were distributed over consecutive sub-networks. As for Bayesian networks, the estimation of parameters is computationally demanding and simplifying the graph's topology based on prior knowledge might be necessary.

A deterministic approach for GRN reconstruction are **Ordinary differential equations** (ODEs), describing rates of changes in e.g. gene expression as a function of the state of (all) other genes in the network. The approaches differ in the basic functional form they use, such as linear functions, power law models or nonlinear functions [Wang and Huang, 2014]. Due to a large number of parameters, ODEs are able to model detailed realistic dynamics, identify temporal patterns of a response and detect causal relationships between genes. The inferred networks are directed and signed, and allow the prediction of their behavior under different conditions, like gene knockout, once the parameters are known. ODE models can be applied to steady-state and time-series data and easily integrate prior knowledge or simultaneously model processes like mRNA degradation. However, solving ODEs requires that the number of experiments exceeds the number of parameters, which is greater than the number of genes in the network. This is normally not the case in biological practice. Assuming that GRNs are unlikely to be fully connected, the number of genes in the model is usually limited or multiple time points are combined to solve the dimensionality problem [Deng et al., 2017].

To infer static or dynamic regulatory networks, different data types can be used. Since transcription is considered the main control mechanism in gene expression, GRN reconstruction usually is based on expression levels [Lappalainen et al., 2013] measured via microarray or RNA-seq, but some methods also can incorporate other omics data like proteomics or metabolomics originating from e.g. mass spectrometry. More complex models include also other sample based measurements like copy number variation (CNV), DNA methylation, somatic mutations or chromatin state measurements. The integration of heterogeneous biological information from multiple omics platforms may enhance the capabilities of GRN inference [Delgado and Gómez-Vela, 2019]. Further, external prior biological knowledge about regulatory effects from databases and the literature can be in-

cluded, like TF bindings sites, binding affinities or miRNA - gene interactions. Different methods for the automatic extraction of relationships between molecular elements from the literature have been developed [Fluck and Hofmann-Apitius, 2014; Habibi et al., 2017; Thomas et al., 2015]. In this thesis, we mainly focus our quantitative analyses on data from steady-state measurements (see Chapter 4), but methods incorporating time-series data exist as well (see Chapter 3).

2.3.3. Challenges

Regulatory mechanisms in living cells are highly complex, posing multiple challenges to their computational analysis. First of all, the quality and performance of inferred regulatory networks is strongly dependent on the amount and quality of available input data, which in turn, even today, are time and cost intensive to obtain, especially in higher organisms. Many large-scale data sets of high throughput experiments have been published and are available in public repositories such as the Gene Expression Omnibus (GEO) [Edgar et al., 2002], the Cancer Genome Atlas (TCGA) [The Cancer Genome Atlas Research Network, 2008] or the Encyclopedia of DNA Elements (ENCODE) [Gerstein et al., 2012]. However, publically available experiments are scattered over many data bases, have various study designs, are based on specific measurement protocols and provide different metadata, making a systematic curation and processing difficult.

Further, the evaluation of computationally predicted GRNs is challenging, as no gold standards for networks of higher organisms exist. Even precise size and density estimations for these networks are lacking [Hart et al., 2006; Stumpf et al., 2008]. In principle, experimental validations of inferred regulatory relationships should be conducted, but those experiments are expensive, especially since a high number of inferred regulatory relations is incorrect. Often, simulated data is therefore used to evaluate GRN reconstruction methods. Regularly, DREAM challenges are organized, which are public competitions to compare network inference algorithms on simulated expression data from simple organisms like E.coli [Marbach et al., 2012]. TF perturbation studies, like knockout or knockdown experiments, where the protein of interest is eliminated or reduced in its amount, can be used to reveal regulatory functions, as changes compared to normal controls are likely to be triggered by the perturbed TF. We used such experiments for our quantitative comparison of different methods estimating regulatory activity in Chapter 4.

Additionally, the number of samples is almost always much smaller than the number of genes considered in a GRN, leading to a dimensionality problem in network inference. To decipher the complex interplay of the interacting entities among an exponential number of possible topologies requires the reduction of the set of potential target genes of a regulator, the inclusion of prior available knowledge or other heuristics. Such priors are used by many methods to set up an initial network structure as baseline for subsequent reconstruction or inference of other parameters and thus reduce the search space (see Section 2.3.4). Knowledge can be included from other biological experiments, databases

or the scientific literature. Even though if the comprehensive integration of heterogeneous biological data into one model is complex, such semi-supervised approaches were shown to outperform unsupervised ones and to be the most successful approaches to GRN inference so far [Pataskar and Tiwari, 2016; Wang and Huang, 2014].

2.3.4. Inference of Regulatory Activity

An important variation of network reconstruction, that partly addresses the dimensionality problem, is the inference of regulatory activity [Wang and Huang, 2014], as studied in this thesis. In this specific case, not the formal structure of the GRN itself is searched, but the states of regulatory elements like transcription factors are inferred, resulting in ranked lists of regulators according to their activity. The activity of regulatory elements can be inferred by combining prior knowledge about potential regulator-gene interactions and data from biological experiments [Brent, 2016]. These network based approaches, combined with an integrative data analysis, can be used for the discovery of biomarkers relevant to diseases or biological processes under investigation. Typical methods rank regulatory features based on their discriminative power comparing different cellular states, like healthy vs diseased conditions [Balwierz et al., 2014; Fröhlich, 2015; Li et al., 2014]. We review some of the these methods in Chapter 3 and quantitatively assess their results in Chapter 4.

By "activity", we refer to a (measurable) effect that a regulator causes by activating or inhibiting a certain target gene in a given context. For example, the reduction of activity in a TF can change the transcription rate of the target gene. Changes in the abundance of the TF protein, its localization, its association with other proteins, or its post-translational modifications may alter a TF's activity [Brent, 2016]. In the case of a change in regulatory activity, a shift of e.g. gene expression levels is expected (see Figure 2.6), and this relationship does not necessarily represent a linear correlation between the activity of a regulator and the mRNA levels of the target genes [Li et al., 2014]. However, the mere differential expression is not a sufficiently good predictor for differential regulator activity, since the observed differential expression pattern is usually a superposition of responses from various influences [Gao et al., 2004]. Furthermore, some genes only change their expression if several TFs are active and interact. Changes of regulatory activity may also result in different states of a biological system without the effect that the target genes or proteins are differentially expressed [Lichtblau et al., 2017; Wu et al., 2016], since a differentially active TF does not necessarily regulate all its target genes [Berchtold et al., 2016].

While established methods to measure mRNA levels of gene expression exist, there are no experimental high-throughput methods determining the activity or inactivity of regulators, like TFs or miRNAs [Berchtold et al., 2016]. Available experimental approaches to infer such activities are ChIP and perturbation studies, like knockout or knockdown experiments. Both techniques can only consider the bindings sites (ChIP) or affected genes (knockout or knockdown) of one or a small number of TFs simultaneously. Further,

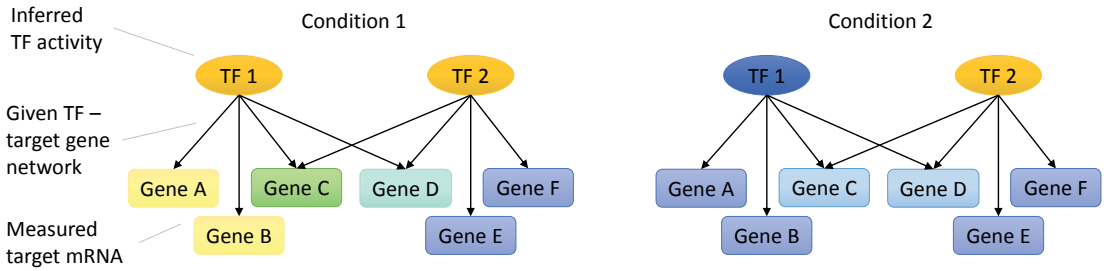


Figure 2.6.: TF activity inference from expression profiles and a TF-target network. High (low) TF activity values and observed mRNA levels are marked in yellow (blue). Adapted from [Brent, 2016].

ChIP experiments cannot distinguish between up- or downregulation of the expressed gene, and detecting the binding of a TF does not necessarily lead to regulation due to post-translational modifications.

Therefore, computational methods have been proposed to determine the activity of TFs. Based on a regulator-gene network, the activity of a regulator can be estimated from the mRNA levels of its direct target genes and other factors like DNA methylation or CNV: Examples for such methods include ISMARA (Integrated System for Motif Activity Response Analysis) [Balwierz et al., 2014], biRte (Bayesian inference of context-specific regulator activities and transcriptional networks) [Fröhlich, 2015], RABIT (Regression Analysis with Background Integration) [Jiang et al., 2015] and RACER (Regression Analysis of Combined Expression Regulation) [Li et al., 2014]. TF activity inference can shed light on unobserved biological processes, including cell cycle [Yang et al., 2005], immune cell differentiation [Yosef et al., 2013] or cancer [Balwierz et al., 2014]. These methods model genome-wide gene expression as sets of (linear) equations over the activity and relationships of transcription factors, genes and other factors and optimize parameters to fit the measured expression intensities. General modeling and optimization approaches are presented in section 2.4 and the specific methods will be described in chapter 3.

2.4. Modeling and Mathematical Optimization

Regulatory activity estimation bases on a vast search space, making optimization algorithms an attractive method [Delgado and Gómez-Vela, 2019]. Theoretically, calculating the activities from samples with thousands of genetic features with an enormous number of possible configurations would require an enormous amount of biological data to ensure the result's reliability [Opgen-Rhein and Strimmer, 2007]. In practice, the number of available expression data is always smaller than the number of investigated genes. Different techniques are applied to face this issue. One widely used approach is feature selection, i.e the choice of a subset of relevant genetic features before the inference step,

removing redundant or irrelevant genes without much loss of information [Bermingham et al., 2015]. Further, the initial choice of a network topology can simplify the inference task. GRNs were shown to be sparse (usually a gene has only a small number of regulators), scale-free (the node degree distribution equals a power law), and modular (consisting of densely connected subsets of nodes that are sparsely linked to the remaining network) [Valverde et al., 2015; Zhang and Zhang, 2013]. These properties can be used to reduce the combinatorial complexity of the inference problem [Babtie et al., 2014]. In this thesis, we focus on methods that restrict the number of potential regulator-gene interactions by integrating prior knowledge about the network's structure. The in Chapter 3 presented and in Chapter 4 evaluated inference strategies include a mathematical optimization step to fit the model to the available data, taking into account prior knowledge and/ or a network template.

2.4.1. Optimization

Optimization comprises an objective function, defining the optimization's goal, and the algorithm itself, actually solving the optimization problem to find a solution that generates the best possible result. In the case of the estimation of regulatory activity, the objective function usually consists of minimizing the absolute sum of errors when comparing measured and predicted expression values or maximizing a likelihood function [Delgado and Gómez-Vela, 2019]. Together with the characterization of particular methods for estimating regulatory activity in Chapter 3, a more detailed description of the applied optimization techniques is provided. The concrete paradigms used for solving the optimization problem include deterministic algorithms, that provide a unique solution when the model input and constraints are fixed, and probabilistic models, where the incorporation of random variables leads to different outputs in each model run and which might be used to estimate probability distributions. A common approach to solve the system of linear equations over the activity and relationships of regulators and genes is linear programming [Bazaraa et al., 2011]. Here, a linear objective function, subject to linear constraints, can be optimized. Stochastic optimization methods generalize deterministic methods, and many different ways to add stochasticity to the same deterministic model frame exist [Fouskakis and Draper, 2002]. Random variables can appear in the formulation of the optimization problem itself or the optimization technique comprises random iterates [Spall, 2005].

The in Chapter 3 presented methods use different models to describe the relationships between measured biological experimental data, the underlying regulatory network and regulatory activity, including linear regression or Bayesian inference. Usually, a regularization term is added, to find a trade-off between the best fit of the model to the data and the solution with a small norm to improve the model's prediction accuracy and interpretability. An example of such a procedure is LASSO (least absolute shrinkage and selection operator) [Tibshirani, 1996], which induces sparsity of the solution by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value. Through the use of the L_1 norm, certain coefficients are set to zero, effectively

choosing a simpler model. A similar idea is incorporated in ridge regression, also known as Tikhonov regularization [Tikhonov, 1963], in which the sum of the squares of the coefficients is forced to be less than a fixed value. This regularization shrinks the size of the coefficients without setting them to zero, not performing variable selection. LASSO can be interpreted as a Bayesian model using Laplacian priors instead of Gaussian priors in the regression framework obtaining point estimates of the regulatory activities and enforcing sparseness of the solution [Li et al., 2014]. Conversely, Bayesian models infer the posterior distribution of the regulator activities by combining the Gaussian likelihood with Gaussian priors for the activities. A method that combines the L_1 and L_2 penalties of LASSO and ridge regression with a mixing parameter is called elastic net [Zou and Hastie, 2005]. This method first applies ridge regression and subsequently commits a LASSO type shrinkage. We will, inter alia, describe the specific optimization and regularization techniques used by different methods for estimating regulatory activity in Chapter 3.

2.4.2. EM Algorithm

Another approach to estimate the parameters of a linear system from observed data is the expectation-maximization (EM) algorithm [Dempster et al., 1977], an optimization technique for solving for maximum-likelihood estimates [Bersanelli et al., 2016]. The principle of the EM algorithm is to iteratively compute maximum likelihood estimates of statistical parameters from data with unobserved variables, called latent variables. After the initialization, each iteration consists of two phases, an expectation (E) step, and a maximization (M) step. During the E step, the expectation of the log-likelihood is evaluated using the current parameter estimates for the latent variables. The M step consists of estimating the parameters maximizing the log-likelihood found on the E step, thus updating the parameter estimates for the next E step [Vaske et al., 2010]. In this thesis (see Chapter 5), the EM algorithm is used for parameter inference where the latent variables correspond to the status of the TF (active or inactive). The procedure provides posterior probabilities of TF activity in a given sample [Picchetti et al., 2015].

The term incomplete data indicates the presence of two sample spaces X and S and a mapping from S to X. The observed data x are realizations from X. The corresponding s in S are not observed directly, but only indirectly through x. Here, we assume that x represents the measured gene expression data, and s are the unobserved TF activity values, which influence the expression of x, but cannot be measured directly. We assume, that each observed gene expression value in a sample has a corresponding unobserved TF activity value. Further, we expect that in each sample group the observations are drawn from specific distributions with corresponding parameters (for example case and control group or knockout and wild type group), and that a further parameter specifies the mixture component. All unknown parameters are collected in the vector θ.

Finding a maximum likelihood solution requires taking the derivatives of the likelihood function with respect to all unknown parameters and simultaneously solving the resulting equations. The EM algorithm solves these two sets of equations numerically and iterates until the results converge. The convergence to a (local) maximum or a saddle point of the likelihood function can be proven [Wu, 1983], but the specific maximum found depends on the starting values.

The likelihood function and its logarithm for n observations can be written as

$$L(\theta; X, S) = P(X, S|\theta) = \prod_{i=1}^{n} P(X_i, S_i|\theta)$$

$$\log(L(\theta; X, S)) = l(\theta; X, S)) = \log(P(X, S|\theta)) = \sum_{i=1}^{n} \log(P(X_i, S_i|\theta).$$

In iteration j with $j = 1...J$ and J the final iteration where convergence is achieved, the E step uses the expected value of $l(\theta; X, S)$ with respect to the current parameters θ^j, defining $Q(\theta|\theta^j)$:

$$Q(\theta|\theta^j) = E(l(\theta; X, S)).$$

Using $E(X) = \sum_{i=1}^{n} x_i \pi_i$ the expectation function for a random variable X with a finite number of outcomes x_i occurring with probabilities π_i respectively, we find

$$Q(\theta|\theta^j) = \sum_{i=1}^{n} \sum_{s} P(S_i = s|X_i = x, \theta^j) \log(P(X_i, S_i = s|\theta)).$$

During the M step, the parameters maximizing Q are found:

$$\theta^{j+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^j).$$

Subsequently, the updated parameters θ^{j+1} are used in a new computation of Q, iterating the E and M step until $Q(\theta|\theta^j) \leq Q(\theta|\theta^{j-1}) + \epsilon$ with a preset threshold ϵ.

For our problem of activity inference, we assume that the gene expression values are drawn from two Gaussian distributions with parameters μ_0, σ_0 for the samples with inactive TFs ($S = 0$), and μ_1, σ_1 for the samples with active TFs ($S = 1$):

$$P(X|S = 0) \sim N(\mu_0, \sigma_0^2), \quad P(X|S = 1) \sim N(\mu_1, \sigma_1^2),$$

with densities

$$f(X_i = x|S_i = s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp - \frac{(x - \mu_s)^2}{2\sigma_s^2}.$$

The mixture of the two Gaussian distributions is given by

$$m(x) = p \cdot f(X_i = x|S_i = 0) + (1 - p) \cdot f(X_i = x|S_i = 1)$$

with p the mixture component reflecting the weighting factor of the sum of both components:

$$p = P(S_i = 0) \qquad \text{and} \qquad 1 - p = P(S_i = 1)$$

where all S_i are considered independent.

Developing the first part of $Q(\theta|\theta^j)$, $\tau_s = P(S_i = s|X_i = x, \theta^j)$ for the activity states $s \in \{0, 1\}$, by using Bayes' theorem, we find for τ_0 and τ_1 in sample i (adapted from [Nuel, 2013]):

$$\begin{aligned}
\tau_0(i) &= P(S_i = 0|X_i = x, \theta^j) \\
&= \frac{P(S_i = 0, X_i = x)}{P(S_i = 0, X_i = x) + P(S_i = 1, X_i = x)} \\
&= \frac{P(S_i = 0) f(X_i = x|S_i = 0)}{P(S_i = 0) f(X_i = x|S_i = 0) + P(S_i = 1) f(X_i = x|S_i = 1)} \\
&= \frac{p \cdot f(X_i = x|S_i = 0)}{p \cdot f(X_i = x|S_i = 0) + (1 - p) \cdot f(X_i = x|S_i = 1)} \quad \text{and} \\
\tau_1(i) &= 1 - P(S_i = 0|X_i = x, \theta^j) \\
&= 1 - \tau_0(i).
\end{aligned}$$

Thus, we can specify

$$\begin{aligned}
Q(\theta|\theta^j) &= \sum_{i=1}^{n} \sum_{s} \tau_s(i) \log(P(X_i, S_i = s|\theta)) \\
&= \sum_{i=1}^{n} \tau_0(i)(\log(p) - log(\sigma_0) - 0.5 \log(2\pi) - (X_i - \mu_0)^2/2\sigma_0^2) \quad + \\
&\qquad \tau_1(i)(\log(1 - p) - log(\sigma_1) - 0.5 \log(2\pi) - (X_i - \mu_1)^2/2\sigma_1^2)
\end{aligned}$$

In the M step, all parameters can be maximized independently, since they all appear in separate linear terms. Considering p we find,

$$\begin{aligned}
\frac{\partial Q(\theta|\theta^j)}{\partial p} = \sum_{i=1}^{n} \frac{\tau_0(i)}{p} - \frac{\tau_1(i)}{1 - p} = 0 \quad \text{thus} \\
p^{j+1} = \sum_{i=1}^{n} \frac{\tau_0(i)}{\tau_0(i) + \tau_1(i)} = \frac{1}{n} \sum_{i=1}^{n} \tau_0(i).
\end{aligned}$$

Analogously, the estimates for μ_s and σ_s can be obtained via

$$\begin{aligned}
\mu_s^{j+1} &= \frac{\sum_{i=1}^{n} \tau_s(i) X_i}{\sum_{i=1}^{n} \tau_s(i)} \quad \text{and} \\
\sigma_s^{j+1} &= \sqrt{\frac{\sum_{i=1}^{n} \tau_1(i)(X_i - \mu_s)^2}{\sum_{i=1}^{n} \tau_1(i)}}.
\end{aligned}$$

Thus, we calculate a new estimate for θ and use it in the next E-step, until convergence of Q. We use the final estimations of $\tau_0(i)$ as probability measure of the inactivity of the TF sample i. Our application of the EM algorithm to the estimation of transcriptional activity in feedback loops is described in more detail in Chapter 5.

3. Computational Methods for Estimating Gene Regulatory Activity

The elucidation of human regulatory relationships is an important research field and many methods attempting to infer the actual regulatory events in a given sample have been proposed, ranging from purely qualitative methods [Liang et al., 1998] over simple statistical approaches [Bansal et al., 2007] to more advanced probabilistic frameworks [Li et al., 2007b]. Early methods were plagued by insufficient data and a general scarcity of background knowledge, which led to rather unstable results [Markowetz and Spang, 2007]. This situation has changed dramatically over the last years, as results of more and more large screens have been made publicly available [Rung and Brazma, 2013] and also the knowledge on principal regulatory relationships has increased [Krämer et al., 2014; Thomas et al., 2015]. This, in turn, has increased the interest in methods which predict genome-wide networks and infer regulatory activity using a systematic, unified, mathematical framework.

Here, we qualitatively review five recent methods for estimating gene regulatory activity with the goal to identify their mutual strengths and weaknesses (see Chapters 4 and 5 for quantitative results). They all assume both the set of regulators (transcription factors or micro RNAs) and the topology of the regulatory network to be given. By combining this background knowledge with specific omics data sets, especially transcriptome data, they try to infer the activity of regulators in a certain experimental condition or disease using optimization of an objective function comparing predicted results with experimental measurements [Ellwanger et al., 2014]. All presented methods are global methods in the sense that they compute activities genome-wide (as much as represented by the underlying network), thus removing the shortcomings of local methods, like ARACNE [Margolin et al., 2006], which ignore cross-talk between sub-models and global effects within samples [Markowetz and Spang, 2007]. The methods predominantly produce a ranked list of regulators, sorted by their activity in a given group of samples. Considering that a multitude of biological influences is ignored during inference, especially kinetic and temporal effects, their goal cannot be to produce precise snapshots of regulatory activity [Budden and Crampin, 2016]. However, several studies reported that such methods can be used to identify biomarkers for specific phenotypes in human cell lines and in vivo samples, for example in innate immunity [Balwierz et al., 2014], ageing related changes [Balwierz et al., 2014] or acute myeloid leukemia [Li et al., 2014].

To emphasize the common ground of these at-first-sight rather different methods, we explain their underlying models by describing differences to a simple framework for

defining the relationships of transcription factors and genes. This framework is presented first; it should be understood as a least common denominator, not as a proper method for network inference by itself. We then describe five recently published methods for genome-wide TF activity estimation as extensions or constraints to this general framework, namely the approach by [Schacht et al., 2014] (estimation of TF activity by the effect on their target genes), RACER [Li et al., 2014], RABIT [Jiang et al., 2015], ISMARA [Balwierz et al., 2014] and biRte [Fröhlich, 2015]. We describe each method in detail and qualitatively compare them with respect to the most important properties, such as the data being used, the method applied for deriving optimized activity values, or the evaluation performed to show effectiveness in the original papers. As baseline, we contrast these more comprehensive methods with the local inference algorithm ARACNE [Margolin et al., 2006], a tool for the de-novo reconstruction of gene regulatory networks. ARACNE requires no background knowledge, but is still rather popular. Key properties of all methods (input, mathematical model, computation, output) are summarized in Table 3.1.

3.1. Mathematical Framework

To combine regulatory networks and quantitative omics data and to thereby deduce regulatory activity, all methods described here use a genome-wide mathematical model. Sample specific gene expression values $g_{i,s}$ for in total n_{genes} genes and $n_{samples}$ samples need to be provided as input. The background regulatory network is represented as a directed graph where the nodes designate regulators and regulated entities (mostly TFs and genes, but also miRNAs, regulatory sites, or TF complexes) and directed edges indicate a regulatory relationship between the two connected nodes, for example the influence of a TF on the expression of a gene.

We will use the variable t for regulators, i for regulated entities, and $b_{t,i}$ for the strength of an edge from a TF or miRNA t to a gene i representing, for instance, a binding affinity. As abstract framework for explaining the different methods we use a simple linear model predicting gene expression $\widehat{g_{i,s}}$ of gene i in sample s in terms of the activity $\beta_{t,s}$ of transcription factor t, that regulates i, in sample s, i.e. considering the binding affinities $b_{t,i}$:

$$\widehat{g_{i,s}} = \sum_{t=1}^{T} \beta_{t,s} b_{t,i}$$

Given this simple model and a set of quantitative measurements of gene expression $g_{i,s}$, the goal of the optimization is to find parameters β such that the sum of squared errors of measured vs predicted gene expression over all genes and samples is minimized using a certain norm, for example the L_2 norm:

$$\min \sum_{i=1}^{n_{genes}} \sum_{s=1}^{n_{samples}} (g_{i,s} - \widehat{g_{i,s}})^2.$$

Method	Input	Model	Computation	Output
Approach by Schacht et al.	- mRNA expression data - TF binding information	Linear model: $\widehat{g_{i,s}} = c + \sum_t \beta_t b_{t,i}(\theta_{a,t} act_{t,s} + \theta_{g,t} g_{t,s})$ with $act_{t,s} = \frac{\sum_i b_{t,i} g_{i,s}}{\sum_i b_{t,i}}$, $\theta_{a,t} + \theta_{g,t} = 1, \theta_{a,t}, \theta_{g,t} \in \{0, 1\}$	- Optimization criterion: Minimize sum of absolute errors - Mixed-integer linear programming - Optimization via Gurobi 5.5	- parameter for each TF: β_t - decision for each TF if $\theta_{a,t}$ or $\theta_{g,t}$ was chosen
RACER	- mRNA expression data - copy number variation - DNA methylation - miRNA expression signals - TF binding information - miRNA target site info (c)	Linear models: 1) $\widehat{g_{i,s}} = c + \theta_{CNV,s} CNV_{i,s} + \theta_{DM,s} DM_{i,s} + \sum_t \beta_{t,s} b_{t,i} + \sum_{mi} \beta_{mi,s} c_{i,mi} miRNA_{mi,s}$ 2) $\widehat{g_{i,s}} = \tilde{c} + \tilde{\theta}_{i,CNV} CNV_{i,s} + \tilde{\theta}_{i,DM} DM_{i,s} + \sum_t \gamma_{i,t} \beta_{t,s} + \sum_{mi} \gamma_{i,mi} \beta_{mi,s}$	- Optimization criterion: Minimize sum of squared errors with $L1$ norm penalty on linear coefficients - Elastic-net regularized generalized linear models and LASSO	1) sample-specific TF and miRNA activities $\beta_{t,s}$ and $\beta_{mi,s}$ 2) TF-gene $\gamma_{i,t}$ and miRNA-gene $\gamma_{i,mi}$ interactions across all samples
RABIT	- differential mRNA expression data - somatic mutations - DNA methylation - copy number variation - TF binding info - recognition motifs for RNA-binding protein (RBP)	Linear model: $\widehat{g_i} = \sum_f \theta_f B_{f,i} + \sum_t \beta_t b_{t,i}$ With B: Background factors (gene CNA, promoter DNA methylation, promoter degree promoter CpG content)	- Frisch-Waugh-Lovell method, select subset of significant TFs via model selection procedure and remove TFs with insignificant correlation across tumors	- regulatory activity score for each TF (t value of linear regression coefficient of t-test)

ISMARA	- gene expression or chromatin state measurements - annotation of promoters (number of predicted sites for motifs) - transcripts and associated promoters - miRNA target site predictions	Linear model: $\widehat{g_{p,s}} = c_p + c_s + \sum_m N_{p,m}\beta_{m,s}$	- Optimization criterion: Minimize sum of errors - Bayesian procedure, ridge regression - Gaussian prior for $\beta_{m,s}$ to avoid overfitting	- inferred motif activity profiles $\beta_{m,s}$ with TFs/ miRNAs binding to motif sites (key regulators) - predicted target promoters, associated transcripts and genes - interaction network of predicted targets and regulators - ontology enrichment
biRte	- mRNA differential expression - miRNA, TF measurements, CNV (optionally) - regulator (R) – target network	Likelihood model: $L_{D,\theta}(R) = p(D\|R,\theta) = \prod_{\hat{D}} p(\hat{D}\|R,\theta) = \prod_{\hat{D}} \prod_c \prod_i p(\hat{D}_{ic}\|R_c,\theta)$	- data specific marginal likelihoods using estimation of hidden state variables with via MCMC - Nested effects model structure learning to reconstruct transcriptional network	- Estimation of active regulators - Estimation of associated transcriptional network
ARACNE	- microarray expression profiles	none	- local estimation of pairwise gene expression profile mutual information	- Reconstruction of gene regulatory network

Table 3.1.: Overview of methods for estimating regulatory activity from transcriptome data comparing input data, modeling, computational aspects and outcome variables. Gene expression data is named g with index i, estimated parameters with β, TF binding information with b, TFs with t, samples with s, miRNAs with mi and model constants with c. Other variables are explained in the text.

Using this model, we assume that the gene expression can be predicted from regulator activities via the linear model, once the regulatory activity is known. In contrast to Figure 2.5, where TFs influence each other, this model, as well as the presented methods in this chapter, ignore TF – TF relations and feedback loops, despite their known importance for gene regulation [Brandman and Meyer, 2008; Sauro, 2017]. Therefore, we later propose a new method for estimating transcriptional activity with a particular focus on the consideration of feedback loops (see Chapter 5). Further, this simple model captures only linear relationships between gene expression and TF activity, ignoring e.g. saturation effects.

3.2. Considered methods

We describe in detail five methods which infer transcription factor activity from omics data sets using a background network of transcription factors and the genes they regulate. These models allow for the application of mathematical optimization to find parameters that minimize the divergence of predicted and measured expression intensities. We further describe the local inference algorithm ARACNE [Margolin et al., 2006], a method for the de-novo reconstruction of gene regulatory networks.

3.2.1. Estimation of TF Activity by the Effect on their Target Genes

The idea of the method by [Schacht et al., 2014] is to use the expression levels of TF's target genes to infer their integrated effect (see Figure 3.1). The method uses expression data and TF binding information as input. The TF – gene network, assembled from different databases, is restricted to genes regulated by more than 10 TFs and TFs with at least 5 target genes. The activity of a TF is modeled linearly by its cumulative effect on its target genes normalized by the sum of target genes or the TF's gene expression level:

$$\widehat{g_{i,s}} = c + \sum_t \beta_t b_{t,i}(\theta_{a,t} act_{t,s} + \theta_{g,t} g_{t,s})$$

where $\widehat{g_{i,s}}$ denotes the predicted gene expression of gene i in sample s, c is an additive offset, β_t describes the estimated activity of TF t and $b_{t,i}$ refers to the underlying strength of the relation between TF t and gene i reflecting the binding affinity. The model is closely related to the general framework mentioned above, only adding a term for the sample specific effect of a TF. The estimated effect of a TF in a certain sample is calculated via the switch-like term in parentheses, where either the activity definition $act_{t,s} = \frac{\sum_i b_{t,i} g_{i,s}}{\sum_i b_{t,i}}$ or the gene expression of the TF itself $g_{t,s}$ is taken into account using the restrictions $\theta_{a,t}, \theta_{g,t} \in \{0, 1\}$ and $\theta_{a,t} + \theta_{g,t} = 1$. This switch term represents a meta-parameter to find the best model and has no biological interpretation. The model outputs an activity value and the information which switch parameter is chosen for each TF of the reduced network.

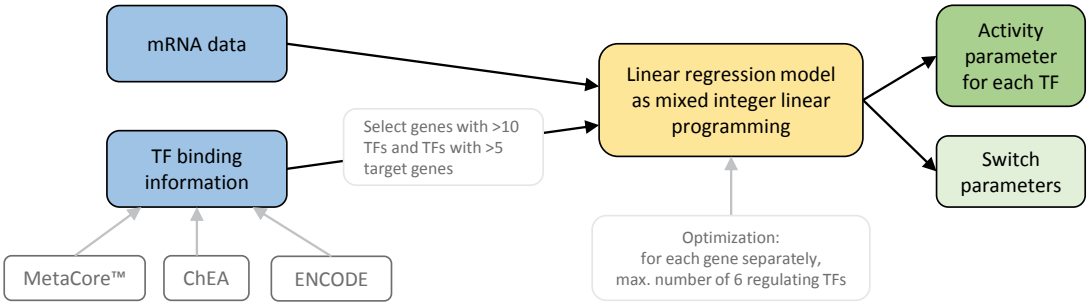


Figure 3.1.: Flow chart of the approach by [Schacht et al., 2014]. The input data sets (marked in blue) are partly filtered and passed to a linear regression model (yellow) which calculates an activity value for each TF (green).

During the optimization, the sum of error terms (absolute value of the difference between predicted and measured gene expression) is minimized which is achieved via mixed-integer linear programming using the Gurobi 5.5 optimizer[1]. The authors of this method state that activity $act_{t,s}$ (see above) was used in 95% of their test cases, but the switch-like combination of both terms yielded still better optimization results. In the paper, the optimization task is greatly simplified as the model is computed for each gene separately and allows only a maximum number of 6 regulating TFs. The TF – gene network indicating the strength of a relation between a TF and a gene is created for 1120 TFs using knowledge from the commercial MetaCore^{TM} database[2], ChEA [Lachmann et al., 2010] and ENCODE [Gerstein et al., 2012]. Due to the restriction of the network mentioned above, the actual model is then based on 521 TFs and 636 target genes only.

Evaluation of the results was performed using expression data from 59 cell lines of the NCI-60 panel [Liu et al., 2010; Shoemaker, 2006] and from melanoma cell lines ("Mannheim cohort") [Hoek et al., 2006]. For each investigated gene, a sample based leave-one-out and 10-fold cross validation of predicted and measured gene expression yielded on average Pearson correlation scores of about 0.6 for both data sets. A gene set enrichment analysis of the target genes for TFs modeled by the activity definition yielded 64 significantly enriched concepts including cell cycle, immune response and cell growth for the data from the NCI-60 panel. Additionally, a t-test was computed between melanoma and other cell lines of the NCI-60 panel to find differentially expressed genes of melanogenesis. For the resulting genes, regulation models were built and used to predict gene expression in the melanoma cell line data set yielding good prediction performances.

[1] http://www.gurobi.com/products/gurobi-optimizer, accessed 09 September 2019
[2] https://portal.genego.com/, accessed 09 September 2019

3.2.2. RACER

RACER (Regression Analysis of Combined Expression Regulation) [Li et al., 2014] aims to integrate generic cell-line data with sample-specific measurements using a two-stage regression (see Figure 3.2). Compared to our general framework, RACER additionally includes miRNA binding information. It assumes a linear combination of the regulatory effects of TFs and miRNAs on mRNA level, which is not further justified. RACER can incorporate a variety of sample specific data including mRNA and miRNA expression values, CNV and DNA methylation. Optimization is applied twice to reduce model complexity, where the method first infers sample-specific TF and miRNA activities and uses these, in a second step, to compute general TF/ miRNA – gene interactions.

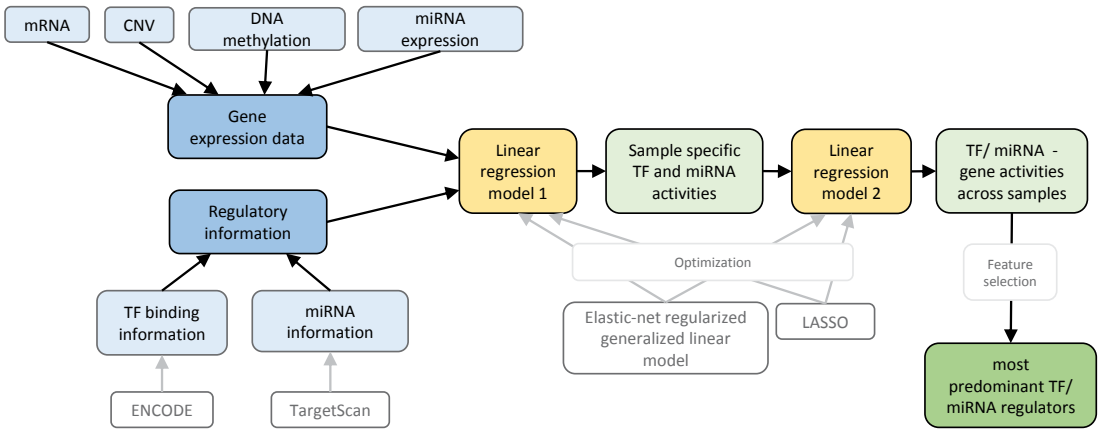


Figure 3.2.: Scheme of RACER method. The input data sets (marked in blue) are passed to a two-step linear regression model (yellow) which calculates sample specific activity values for each regulator and determines the most predominant regulators (green).

In the first regression step, mRNA, miRNA, CNV and DNA methylation data are used to obtain the sample specific activities:

$$\widehat{g_{i,s}} = c + \theta_{CNV,s} CNV_{i,s} + \theta_{DM,s} DM_{i,s} + \sum_{t} \beta_{t,s} b_{t,i} + \sum_{mi} \beta_{mi,s} c_{i,mi} miRNA_{mi,s}$$

where $\widehat{g_{i,s}}$ denotes the predicted gene expression of gene i in sample s, c is an intercept, $\beta_{t,s}$ describes the estimated activity of TF t in sample s and $b_{t,i}$ is the TF – gene binding score for TF t and gene i. The parameter $\beta_{mi,s}$ stands for the estimated activity of miRNA mi in sample s and is multiplied by $c_{i,mi}$, the number of conserved target sites on 3'UTR of the target gene i for miRNA mi, and by the expression level of miRNA mi in sample s. $\theta_{CNV,s}$ (respectively $\theta_{DM,s}$) are the regression parameters for CNV signals $CNV_{i,s}$ (respectively DNA methylation data $DM_{i,s}$).

Using $\beta_{t,s}$ and $\beta_{mi,s}$ from the first regression step, TF – gene and miRNA – gene interactions across all samples are calculated in a second model:

$$\widehat{g_{i,s}} = \tilde{c} + \tilde{\theta}_{i,CNV} CNV_{i,s} + \tilde{\theta}_{i,DM} DM_{i,s} + \sum_{t} \gamma_{i,t} \beta_{t,s} + \sum_{mi} \gamma_{i,mi} \beta_{mi,s}$$

where the sums apply only to a number of selected TFs and miRNAs with nonzero binding signals $b_{t,i} > 0$ and number of conserved target sites $c_{i,mi} > 0$. After the optimization, $\gamma_{i,t}$ and $\gamma_{i,mi}$ indicate the strength of a TF/ miRNA – gene relationship across all samples. To obtain robust estimates, $\gamma_{i,mi}$ is additionally weighted by the averaged activities of the miRNA.

In each of the two regression steps, the optimization criterion is to minimize the sum of squared errors with L_1 penalty on the linear coefficients to induce a sparse solution and to set irrelevant parameters to zero after the fitting. This sparse LASSO solution is obtained through elastic-net regularized generalized linear models [Zou and Hastie, 2005]. A supplementary feature selection comparing the full model to a restricted model leaving one TF or miRNA out provides the most predominant TF/ miRNA regulators. TF binding scores are collected from the generic cell line of erythroleukemia cells K562 from ENCODE for 97 TFs and 16653 genes. Further, the number of conserved target sites on 3'UTR is taken from sequence-based information from TargetScan for 470 miRNAs and 16653 genes. The RACER method is implemented in R and publicly available[3].

The method was evaluated using expression data from an acute myeloid leukemia (AML) data set from TCGA with 173 samples [The Cancer Genome Atlas Research Network, 2013] via a sample based 10-fold cross validation on the prediction of gene expression. To assess the quality of predictions, the Spearman rank correlation was calculated resulting in a reassuring value of approximately 0.6. Further, the full model was compared to models excluding one type of the input variables. The full model performed best and a substantial reduction of Spearman correlation was observed by omitting TF regulation (20%) or DNA methylation (5%). RACER also performed with competitive accuracy in predicting known miRNA – mRNA and TF – gene relationships compared to other methods like GenMiR++ [Huang et al., 2007] or ENCODE TF binding scores [Gerstein et al., 2012] using e.g. validated interactions from the MirTarBase [Hsu et al., 2011] and knockdown studies. The feature selection procedure revealed 18 predominant transcriptional regulators in the AML data set. Using their associated targets, a functional enrichment analysis showed that DNA repair and the tumor necrosis factor pathway were enriched. When applying this panel to cluster patients at different cytogenetic risks, the clustering pattern of the regulatory activities was largely consistent with the risk groups. Further, a literature survey on AML showed that many TF regulators among the top predictions had a role in leukemogenesis.

[3] http://www.cs.utoronto.ca/~yueli/racer.html, accessed 09 September 2019

3.2.3. RABIT

Regression Analysis with Background Integration (RABIT) [Jiang et al., 2015] is a method for finding expression regulators in cancer by a large scale analysis across diverse cancer types. It integrates TF binding information with tumor profiling data to search for TFs driving tumor-specific gene expression patterns (see Figure 3.3). It can be applied to predict cancer-associated RNA-binding protein (RBP) motifs which are key components in the determination of miRNA function [van Kouwenhove et al., 2011].

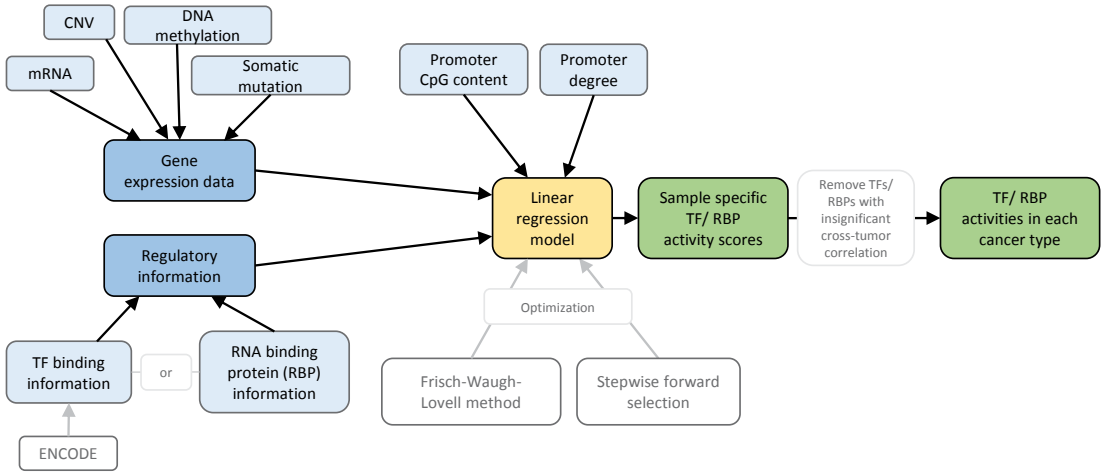


Figure 3.3.: Flow chart of RABIT method. The input data sets (marked in blue) are passed to a linear regression model (yellow) which calculates sample specific activity values for each regulator and determines general regulatory activities (green).

Extending our general framework, RABIT can, like RACER, use CNV and DNA methylation data and additionally integrates promoter CpG content and promoter degree information (total number of ChIP-seq peaks near the gene transcription start site). RABIT takes RBP or TF binding information as regulatory input. The computational model consists of three steps (see Figure 3.3). First, RABIT tests in each tumor whether the target genes, identified by the BETA method [Wang et al., 2013], show differential expression compared to the normal controls including a control for background effects from CNVs, promoter DNA methylation, promoter CpG content and promoter degree:

$$\widehat{g_i} = \sum_f \theta_f B_{f,i} + \sum_t \beta_t b_{t,i}$$

where $\widehat{g_i}$ represents the predicted differential gene expression between tumor and normal samples in gene i, B includes values of the f different background factors for gene i, b contains RBP or TF binding information and θ and β are the respective regression parameter vectors. The regression coefficients β are estimated by minimizing the squared difference between measured and predicted gene expression.

The regulatory activity score is defined for each TF/ RBP as the t-value of the linear regression coefficient t-test and calculated via the coefficient divided by the standard error. If multiple profiles exist for the same TF from different conditions or cell lines, the profile with the highest absolute value of TF regulatory activity score is selected. In a second step, a stepwise forward selection is applied to find a subset of TFs among those screened in step one optimizing the model error. Lastly, TFs with insignificant cross-tumor correlation are removed from the results.

Computationally, the regression coefficients are calculated via the Frisch-Waugh-Lovell method [Frisch and Waugh, 1933]. TF binding information is taken from 686 TF ChIP-seq profiles from ENCODE representing 150 TFs and 90 cell types. Additionally, recognition motifs for 133 RBPs and their putative targets are collected by searching recognition motifs over the 3'UTR regions [Ray et al., 2013]. An implementation of the RABIT method can be downloaded[4].

RABIT was applied to 7484 tumor profiles of 18 cancer types from TCGA using gene expression, somatic mutation, CNV and DNA methylation data. To systematically assess the results, the cancer relevance level of a TF was calculated as percentage of tumors with the TF target genes differentially regulated (averaged across all TCGA cancer types). A comparison to cancer gene databases, i.e. the NCI cancer gene index project [National Cancer Institute Wiki, 2014], the Bushman Laboratory cancer driver gene list [Sadelain et al., 2012; Vogelstein et al., 2013], the COSMIC somatic mutation catalog [Futreal et al., 2004] and the CCGD mouse cancer driver genes [Abbott et al., 2015], showed a consistent picture. Further, RABIT's performance was compared to other regression models like LAR or LASSO where RABIT had the best results when classifying all TFs into three categories by NCI cancer index and achieved better cross-validation error and shorter running time. The regulatory activity of RBPs showed that some alternative splicing factors could affect tumor-specific gene expression by binding to target gene 3'UTR regions.

3.2.4. ISMARA

In contrast to the previous three methods and to our general framework which directly scores TFs or other regulators, ISMARA (Integrated System for Motif Activity Response Analysis) [Balwierz et al., 2014] infers the activity of regulatory motifs (short nucleotide sequences) and thereby only indirectly deduces the effects of TFs and miRNAs (see Figure 3.4). ISMARA is a web service where no parameter settings or specific processing of the input data, gene expression or ChIP-seq data are necessary. It can also be used to calculate regulatory activity differences between samples and consider replicates or data from time series.

[4] http://rabit.dfci.harvard.edu/download, accessed 09 September 2019

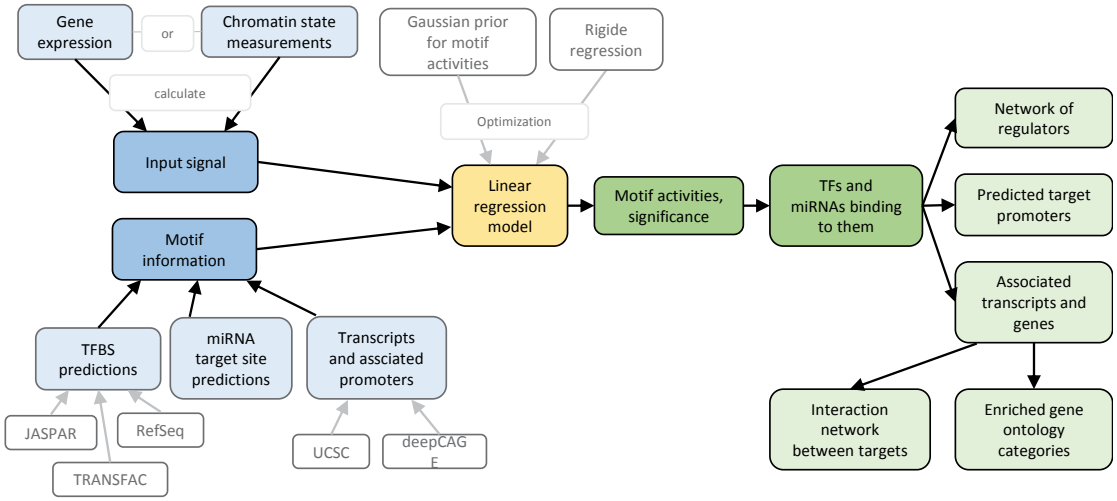


Figure 3.4.: ISMARA model scheme. The input data sets (marked in blue) are passed to a linear regression model (yellow) which calculates motif activities and determines associated regulators (green).

ISMARA takes sample specific measurements and information about regulatory motifs for TFs and miRNAs into account. Based on the input of gene expression data or chromatin state measurements, an input signal is calculated for each promoter in each sample. The input signals are modeled linearly in terms of the binding site predictions and unknown motif activities:

$$\widehat{g_{p,s}} = c_p + c_s + \sum_m N_{p,m}\beta_{m,s}$$

where $\widehat{g_{p,s}}$ refers to the input signal for a promoter p in sample s, c_p and c_s are intercepts for each promoter and sample, $N_{p,m}$ summarizes the TF/ miRNA binding site predictions (sum of the posterior probabilities of all predicted TF/ miRNA binding sites for motif m in promoter p) and $\beta_{m,s}$ stands for the estimated motif activities. Like in the other presented methods, the optimization criterion is to minimize the sum of squared error terms between predicted and measured gene expression.

Primarily, ISMARA provides the inferred motif activity profiles $\beta_{m,s}$ sorted by significance and a set of TFs and miRNAs that bind to these motifs representing the key regulators. Further, a list containing their predicted target promoters, associated transcripts and genes, a network of known interaction between these targets and a list of enriched gene ontology categories is displayed. ISMARA is available as a web service[5] only.

[5]http://ismara.unibas.ch, accessed 09 September 2019

ISMARA employs a Bayesian procedure with a Gaussian likelihood model and a Gaussian prior distribution for $\beta_{m,s}$ to avoid overfitting. Information about regulatory motifs is provided via the annotation of promoters based on deep sequencing data of transcription start sites. To obtain a set of promoters and their associated transcripts, the 5' ends of mRNA mappings from UCSC genome database are clustered with the promoters. TF binding site predictions in the proximal promoter region are collected using 190 position weight matrices representing 350 TFs from JASPAR, TRANSFAC, motifs from the literature and their own analyses of ChIP-seq and ChIP-chip data. Additionally, miRNA target sites for about 100 seed families are annotated in the 3'UTRs of transcripts associated with each promoter.

For evaluation, the original paper applied ISMARA to data from well-studied systems and results were compared to the literature. Inferred motif activities were highly reproducible and even more robust than the expression profiles from which motif activities were derived. When comparing samples from 16 human cell types (GEO accession number GSE30611) from younger and older donors, ISMARA was able to identify a key regulator of aging-related changes in expression of lysosomal genes. A joint analysis of the human GNF atlas of 79 tissues and cell lines [Su et al., 2004] and the NCI-60 reference cancer cell lines [Ross et al., 2000] revealed that many of the top dysregulated motifs were well-known in cancer biology like HIF1A and has-miR-205 miRNA. They also suggested novel predictions for regulating TFs in innate immunity, mucociliary differentiation and cancer.

3.2.5. BiRte

BiRte (Bayesian inference of context-specific regulator activities and transcriptional networks) [Fröhlich, 2015] takes a mathematically different approach compared to the methods described before, integrating TF/ miRNA target gene predictions with sample specific expression data into a joint probabilistic framework (see Figure 3.5). Compared to our general scheme of a TF – gene network (Figure 2.5), biRte first uses the TF/miRNA – gene network without the interactions between regulators themselves (e.g TF-TF interactions) to estimate regulatory activities. BiRte infers the network between the regulators in a second step.

BiRte takes as input differential gene expression data (mRNA), an underlying regulatory network including TF/ miRNA – target gene binding information and optionally CNV data, miRNA and TF expression measurements. BiRte defines a likelihood model for the set of active TFs/ miRNAs (called regulators R which can be seen as hidden variables) based on the entire gene expression data D and certain model parameters θ:

$$L_{D,\theta}(R) = p(D|R,\theta) = \prod_{\hat{D}} p(\hat{D}|R,\theta) = \prod_{\hat{D}} \prod_{c} \prod_{i} p(\widehat{D_{ic}}|R_c,\theta)$$

Here, $\hat{D}$ represents the set of all available experimental data including mRNA, CNV, miRNA and TF expression data and $\widehat{D_{ic}}$ refers to the i^{th} feature measured under exper-

imental condition c. The condition specific hidden state variables R_c are estimated with help of the Markov Chain Monte Carlo (MCMC) method where a regulator can switch from an active to an inactive state (switch) or an inactive and an active regulator exchange their activity states (swap). Thereby, the posterior probability for each regulator and condition to influence the expression of its target genes is estimated. Simultaneously, a variable selection procedure is applied to achieve sparsity of the model. The optimization goal is not, as one would expect, to return the configuration with highest posterior probability among all sampled ones but to take marginal selection frequencies during sampling into account and filter those above a defined cutoff. After the determination of active regulators, the associated transcriptional network between active regulators is estimated based on observed nested subsets of differentially expressed target genes.

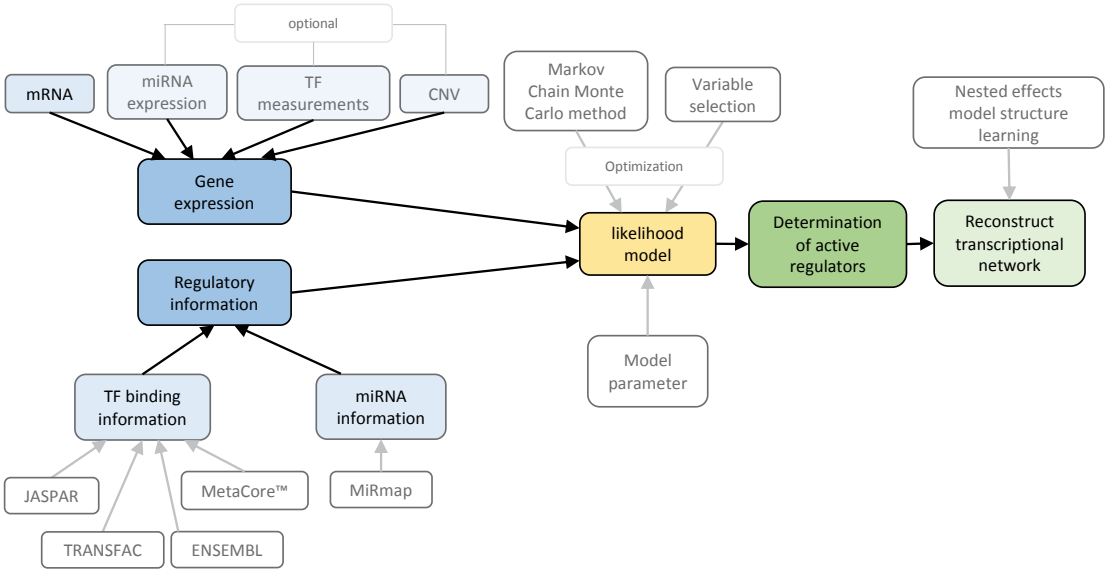


Figure 3.5.: Scheme of biRte method. The input data sets (marked in blue) are passed to a likelihood model (yellow) which determines active regulators (green).

In the implementation, the stochastic sampling scheme based on MCMC allows swap operations only when regulators show a significant overlap of regulated targets. The variable selection procedure is implemented via a spike and slab prior [George and Mculloch, 1997] which can integrate prior knowledge about the activity of regulators. To infer the associated transcriptional network, Nested Effects Model (NEM) [Markowetz et al., 2007] structure learning is applied. An input miRNA – gene network is constructed based on MiRmap [Vejnar and Zdobnov, 2012] for 356 miRNAs. A TF – target gene network with 344 TFs is compiled by computing TF binding affinities to promoter sequences according to the TRAP model [Roider et al., 2007] using data from ENSEMBL, TRANSFAC, JASPAR and MetaCore^{TM}. An implementation of biRte is available for R on Bioconductor[6].

[6]https://bioconductor.org/packages/release/bioc/html/birte.html, accessed 10 September 2019

Several simulations were conducted to study model behavior. On the basis of a human regulatory sub-network and accordingly simulated expression data of 900 target genes, biRte was compared to BIRTA [Zacher et al., 2012], GEMULA [Geeven et al., 2012] and a hypergeometric test and further to other network reconstruction algorithms like ARACNE [Margolin et al., 2006], GENIE3 [Huynh-Thu et al., 2010] and GeneNet [Opgen-Rhein and Strimmer, 2007]. BiRte performed best in regulator activity predictions including a favorable computation time and was robust against false positive and false negative target gene predictions. Additionally, biRte was applied to an E.coli growth control and to a prostate cancer data set including 44 normal and 47 cancer samples from GEO (GSE29079) with corresponding array data from 464 human miRNAs (GSE54516) and the results showed a principal agreement with the biological literature.

3.2.6. ARACNE

We compare ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [Margolin et al., 2006] as an established, yet local, tool for the reconstruction of gene regulatory networks to the previous five recent genome-wide approaches. The algorithm is background knowledge-free and identifies transcriptional interactions based on mutual information including non-linear and non-monotonic relationships and distinguishes between direct and indirect relationships (see Figure 3.6). Networks obtained with ARACNE can be used to calculate regulatory activity using the aforementioned methods. ARACNE is available via a free online tool[7] or as an R package in the `minet` library [Meyer et al., 2008].

ARACNE uses as input only microarray expression profiles and estimates candidate interactions by calculating the pairwise gene expression profile mutual information I defined as

$$I(g_i, g_j) = I_{i,j} = S(g_i) + S(g_j) - S(g_i, g_j)$$

where S denotes the entropy. $I_{i,j}$ measures the relatedness of genes g_i and g_j and equals zero if both are independent. In a second step, the mutual information values are filtered using a threshold depending on the distribution of all mutual information values in random permutations of the original data set. Indirect interactions are then removed.

Computationally, a Gaussian kernel operator is used to calculate mutual information scores. In a subsequent step, the data processing inequality (DPI) [Cover and Thomas, 1991] is applied to remove probably indirect candidate interactions. The DPI states that if the genes g_i and g_k interact only through a third gene g_j, then

$$I(g_i, g_k) \leq \min(I(g_i, g_j), I(g_j, g_k))$$

Thus, the least of the three mutual information scores can come from indirect interactions only [Margolin et al., 2006].

[7] http://califano.c2b2.columbia.edu/aracne, accessed 09 September 2019

ARACNE's performance was evaluated on the reconstruction of realistic synthetic data sets [Mendes et al., 2003] and on an expression profile data set consisting of about 340 B lymphocytes derived from normal, tumor-related and experimentally manipulated populations [Klein et al., 2001]. Regarding the synthetic networks, ARACNE had consistently better precision and recall values compared to the two other algorithms, Relevance Networks [Butte and Kohane, 2013] and Bayesian networks [Hartemink et al., 2001], and reached very good precision at significant recall levels. It recovers far more true connections and fewer false connections than the other methods with better performance on tree-like topologies compared to scale-free topologies. A reconstructed B-cell specific regulatory network was found to be highly enriched in known c-MYC targets where about 50% of the predicted genes to be first neighbors were reported in the literature.

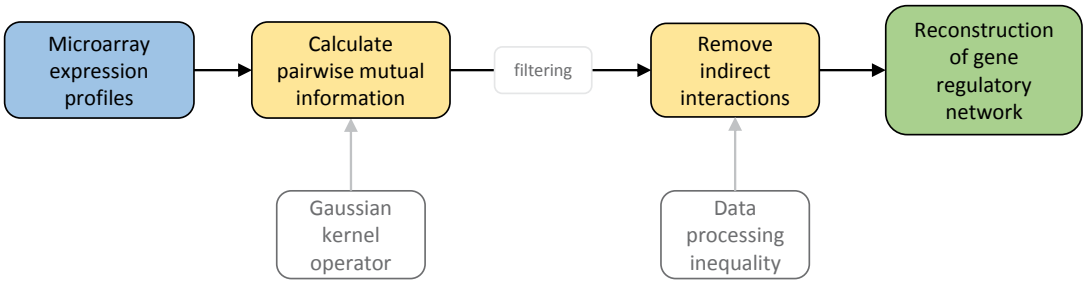


Figure 3.6.: ARACNE flow chart. The input data set (marked in blue) is used to calculate pairwise mutual information where indirect interactions are removed (yellow) and which allow a reconstruction of the gene regulatory network (green).

3.3. Comparison

We described five recent methods for the genome-wide inference of regulatory activity, namely the approach by [Schacht et al., 2014], RACER, RABIT, ISMARA, and biRte. They all assume the topology of the regulatory network to be known, cast activity estimation as an optimization problem regarding the difference between predicted and measured values, take different types of sample specific omics data into account, and eventually produce a list of transcription factors or miRNAs, ranked by their estimated activities in the samples under study. As a baseline, we also included ARACNE which is background knowledge-free and uses only local dependency measures to reconstruct a regulatory network and indirectly infer activities. All of the presented methods essentially follow the same goal, i.e., accurate ranking of regulatory activity, but differ in the types of measurements being integrated, the background knowledge necessary

for their application, the complexity and refinement of the underlying model of gene regulation, and the concrete paradigm used for solving the optimization problem. The methods, except for the approach by [Schacht et al., 2014], are available online via a downloadable implementation (RABIT, ARACNE), a web service (ISMARA), and/or an R package (RACER, biRte, ARACNE) providing an operable solution for the interested user. Whereas an overview of the main features of each method can be found in Table 3.1, we now compare the algorithms regarding their general properties in a descriptive way.

3.3.1. Experimental Data Types

The methods differ in the types of measurements being integrated, which corresponds to the level of detail of their model of gene regulations. All six methods use mRNA as input. RACER, RABIT and biRte can also integrate CNV, DNA methylation, TF/ miRNA expression data, or somatic mutations. ISMARA calculates an input signal from microarray, RNA-Seq, or ChIP-seq data. Additionally, all presented methods, except ARACNE, use prior knowledge about the underlying regulatory network. These networks are extracted from different data sources and preprocessed in different manners. All methods, except ARACNE, require at least knowledge about TF – gene relationships, yet RACER, biRte and ISMARA also incorporate information about miRNAs. When using RABIT, the user can choose whether to provide TF or RNA binding protein information.

The approach of [Schacht et al., 2014] and biRte extract regulatory information partly from the commercial MetaCore^{TM} database, whereas the other methods use only publicly available databases, like ENCODE, JASPAR or TRANSFAC. The networks which are used for the evaluations published in the respective papers are publicly available for the case of RACER (network for 16653 genes, 97 TFs and 470 miRNAs), RABIT (predicted binding scores of 63 RBP motifs and 17463 genes) and biRte (network for E.coli including 160 TFs). Neither [Schacht et al., 2014] nor ISMARA make this data available.

3.3.2. Mathematical Models

The methods use different mathematical models to infer regulatory activity. The method by [Schacht et al., 2014] is the most closely associated one compared to our general framework. RACER and RABIT can be seen as extensions of the approach by [Schacht et al., 2014] since they essentially use the same model structure but incorporate more input data types and more classes of regulatory information. Also chronologically, [Schacht et al., 2014] published their method first, followed by RACER and RABIT. The approach by [Schacht et al., 2014], RACER, RABIT and ISMARA use linear regression whereas biRte applies a probabilistic framework. RACER applies a two-stage regression to infer regulatory activity. ARACNE, as a local method, is based on mutual information. Also, the method by [Schacht et al., 2014] resolves the optimization problem not globally and restricts the number of regulating TFs per gene. The other methods model activities genome-wide, as much as represented by the underlying network.

3.3.3. Optimization Frameworks

For assessing regulator activities, [Schacht et al., 2014], RACER, RABIT and ISMARA minimize the sum of error terms between measured and predicted gene expression. However, the methods use rather different algorithms for solving the resulting optimization problem, and also apply different constraints to achieve model sparsity, robustness of inference, and feature selection. In the approach by [Schacht et al., 2014], the regression model is computed for each gene separately and allows only a maximum number of six regulating TFs. RACER uses a LASSO approach, while ISMARA follows a Bayesian model that infers regulator activities as posterior distributions. LASSO can be interpreted as a Bayesian model using Laplacian priors instead of Gaussian priors in the regression framework. Thus, point estimates of the regulatory activities are obtained and sparseness of the solution is enforced [Li et al., 2014]. In contrast, biRte uses a likelihood model with a spike and slab prior to induce model sparsity. This approach implements a selective shrinkage of model coefficients such that estimates are less biased compared to a LASSO prior [Hernández-Lobato et al., 2010]. With the help of the spike and slab prior, sparsity can be controlled in a variable dependent manner allowing the inclusion of prior belief in the activity of each regulator [Fröhlich, 2015].

3.3.4. Outputs

[Schacht et al., 2014] and biRte determine activity of regulators over all samples at once, whereas RACER and biRte first infer sample-specific activities which are combined to cross-tumor activities in a second optimization step. In contrast, ISMARA in first place infers motifs activity; these activities are used to deduce the effects of TFs and miRNAs by their motif binding profiles. ISMARA primarily provides sample specific TF and miRNA activity but also offers an option to group samples and com-pare average regulatory activity between different conditions. Like biRte and ARACNE, it also infers the network of the regulators themselves.

3.3.5. Evaluations

The type and extent of evaluation performed for the different methods vary greatly. They range from direct application to biological problems over the comparison of results to the biological literature to simulation studies. All methods published evaluations results on publicly available data sets, e.g. from the National Cancer Institute, TCGA or GEO, but unfortunately address different tissues and cancer types. Sample-based cross-validation is applied in the work by [Schacht et al., 2014], RACER, RABIT and ISMARA. The first two of these methods use correlation coefficients between measured and predicted gene expression for assessing prediction quality. The authors of RACER, RABIT and biRte compare their results to the outcome of other algorithms and to those of restricted models, for example excluding one type of the input variables. All methods search the literature to compare their predictions to previously published studies on the respective biological question. Overall, ISMARA provides the most extensive biological evaluation using a battery of relevant use cases, whereas biRte excels in systematic

simulation studies. Sadly, there are very few works which compare any of the methods presented on the same problem; the only result we are aware of compared ARACNE and biRte regarding their performance in network reconstruction on simulated data, in which biRte attained higher robustness against false positive and false negative target gene predictions [Fröhlich, 2015].

3.4. Discussion

3.4.1. Background Networks

A crucial input to the models is the underlying regulatory network which is needed to reduce the search space for actual regulatory activity. However, the construction of comprehensive TF/ miRNA – gene regulatory networks is difficult for various reasons. Firstly, a comprehensive characterization of the human regulatory repertoire is lacking since only about half of the estimated 1,500 - 2,000 TFs in the mammalian genome is known [Vaquerizas et al., 2009]. ChIP experiments, prone to a high false positive rate [Pickrell et al., 2011], were used to identify TF binding patterns but each assay is limited to the detection of one TF in one condition and therefore TF binding has not been characterized for many TFs in most cell types. Further, the local proximity of a binding site to the transcriptional start site of a gene does not automatically implicate transcriptional regulation. With regard to post-transcriptional regulators, the functions for only a few of the around 1,200 different miRNAs have been experimentally determined and current data on miRNA targets is mostly based on computational predictions [Rajewsky, 2006]. Generally, the knowledge about TF and miRNA binding is scattered over the biological literature and different, partly commercial, databases, impeding the construction of comprehensive networks [Thomas et al., 2015].

Any comparative evaluation of the methods presented here would have to must make sure that the same background network is used for each computation. Besides, studies on the impact of network incompleteness or different error rates in networks would be important to assess the ability of the methods to cope with such common problems. Simulation studies will be vital in this regard.

3.4.2. Biological Networks as Di-Graphs

The modeling of regulatory networks as graphs (see chapter 2), as used in all presented methods, is perhaps not the optimal representation for the underlying biological regulatory processes. A graph cannot easily account for important effects such as TF complex formation and temporal and spatial synchronization of activities. Furthermore, TF binding is affected by chromatin state and the impact of post-translational modifications on transcriptional activity which are difficult to include in a graph view on regulation. The model's dependence on the topological structure and the robustness to changes in the underlying network have not been evaluated or discussed in any of the presented methods even if these issues are known to have a severe influence in network analysis [Babtie et al., 2014].

3.4.3. Mathematical Model

Linear models, widely spread in different fields of science, provide a simple and easily understandable design. However, linear models over-simplify the underlying biological processes. Nonlinear behavior, e.g. saturation effects, cannot be represented, as well as feedback loops (see Chapter 5). Considering that the number of available samples is typically relatively small, the incorporation of many different data types and according parameters into the model could result in excessively complex designs prone to overfitting, but this issue lacks general awareness. Only two of the presented methods incorporate parameter priors (ISMARA and biRte), and only two apply cross validation techniques to estimate prediction performance (method by [Schacht et al., 2014] and RACER). Further, the effect of temporal buffering between TF binding and the actual effect on gene expression is not included in any of the methods.

3.4.4. Comparability

All methods produce a ranked list of regulators. Rating these results across different methods, even when applied on the same data set and using the same background network, is difficult since no generally accepted benchmarks are available. Therefore, there currently is no objective measure to designate a best method. The closest comparable evaluation effort we are aware of is implemented in the "DREAM5 – Network Inference" challenge [Marbach et al., 2012], which targets gene regulatory network reconstruction. The invited participants reverse-engineered a network from gene expression data, including a simulated network, and evaluated the results on a subset of known interactions or the known network for the in-silico case in the 2010 edition. The approach of GENIE3 [Huynh-Thu et al., 2010] which trains a random forest to predict target gene expression performed best and the integration of predictions from multiple inference methods showed robust and high performance across diverse data sets. However, an extensive competitive evaluation to determine active regulators based on a given regulatory network has, to the best of our knowledge, not been carried out yet.

3.4.5. Latest Research

The previous analyses in this chapter, comparing different methods for estimating regulatory activity, are based on our publication in BMC Systems Biology [Trescher et al., 2017]. Here, we will give a brief update on newly developed methods and the current status of this research field.

Several new techniques and extensions of existing methods for estimating transcriptional activity have been proposed since then. For example, [Alvarez et al., 2016] developed VIPER (virtual inference of protein activity by enriched regulon analysis), extending a TF estimation method based on a probabilistic framework to the application to single sample expression profiles. They used ARACNE to construct a TF-gene network and detected the maximum information path targets. By computing the enrichment

of a protein's transcriptional targets in differentially expressed genes via analytic rank-based enrichment analysis (areA), a statistical analysis based on the mean of ranks, they inferred differential protein activity as the normalized enrichment score computed by areA. In vitro assays could confirm that protein activity inferred by VIPER outperformed mutational analysis in predicting sensitivity to targeted inhibitors. VIPER is available as an R package on Bioconductor and requires a gene expression signature and an appropriate cell context-specific regulatory network as input. [Garcia-Alonso et al., 2019] applied VIPER to assess drug sensitivity in cancer by analyzing the transcriptional dysregulation in cancer cell lines and patient tumors and published their results as DoRothEA (Discriminant Regulon Expression Analysis), a database with candidate TF-drug interactions in cancer.
A different approach, which is based on network analysis, is LEAN (local enrichment analysis) [Gwinner et al., 2017]. It uses a local subnetwork model and genome-wide omics data sets to identify statistically dysregulated subnetworks and directly suggests single genes for follow-up experiments. LEAN is also available as an R package (`LEANR`), and takes a list of measures of statistical significance for some or all genes and an interaction network as input.
Further, [Xi et al., 2018] use matrix factorization to discover driver genes from mutation data using interaction networks and mRNA expression data. Their method outperformed existing network-based methods and detected new driver genes in cancer. Their implementation is available on GitHub[8] and takes somatic mutations of patients across multiple cancer types, mRNA expression data and an interaction network as input to compute driver gene candidates.
Two methods apply network component analysis (NCA) to quantify transcription factor activities: Local NCA (LNCA) [Shi et al., 2017] and sparseNCA [Noor et al., 2018]. LNCA evaluates the local similarities of regulatory variations by integrating the expression sets and prior TF-gene regulatory knowledge. LNCA was implemented as a Matlab package, but unfortunately not available online at the time of our search. SparseNCA incorporates the effect of incompleteness of the underlying network in the estimation of TF activity and can be downloaded as a C++ implementation[9].
A different approach was published by [Martignetti et al., 2016], ROMA (Representation and quantification of Module Activities), which bases activity quantification on the simplest uni-factor linear model of gene regulation that approximates the expression data of a gene set by its first principal component. ROMA can be downloaded[10] as Java program and computes the activity of gene sets based on gene expression data and gene set definitions.

Further, methods evaluating existing regulatory activity predictions have been proposed. [Sikdar and Datta, 2017] extended the search for active TFs and published a method to identify master regulators, i.e. TFs that control most of the regulatory activities of other TFs. They evaluate the concordance of two ranked TF lists using a

[8]https://github.com/USTC-HIlab/RS-ExpNet-CRNMF, accessed 10 September 2019
[9]https://sites.google.com/site/aminanoor/softwares, accessed 10 September 2019
[10]https://github.com/sysbio-curie/Roma, accessed 10 September 2019

statistical measure, the connectivity score, estimating the change in connectivity of a TF with the genes in different sample groups. Another measure for the systematic evaluation of inferred activity changes was proposed by [Berchtold et al., 2016], called i-score (inconsistency score), quantifying how many genes could not be explained by the set of activity changes of TFs. They observed, that many published methods (like ISMARA) yielded a high number of unexplained target genes (i.e. large i-scores), indicating that currently available regulatory networks are far from being complete. They also developed the theoretical minimum of the score given the expression data and the gene regulatory network, which can be used to evaluate the results of different networks.

3.5. Conclusion

Despite their often rather involved procedures and models, none of the presented methods adequately reflects the biological reality of regulatory activity in cells. A specific disease phenotype is rarely caused by a single gene but rather a product of the interplay of genetic variability, epigenetic modifications and post-transcriptional regulation mechanisms [Davidsen et al., 2016]. The presented methods ignore a multitude of such factors like the effects of chromatin state and alternative splicing, nonlinear relationships between regulatory activity and gene expression, or kinetic and temporal effects. Furthermore, TFs themselves regulate the expression of other TFs forming feedback loops which are not considered in any of the presented methods. We therefore propose a new method for estimating transcriptional activity with a particular focus on the consideration of feedback loops (see Chapter 5). Nevertheless, the methods apparently are able to detect strong signals and are thus valuable tools for identifying biomarkers for specific phenotypes.

4. Evaluation of Methods Scoring Regulatory Activity

As described in the previous chapter, several algorithms have been presented to model genome-wide gene expression and regulation via the activity and relationships of transcription factors. These methods apply mathematical optimization to find parameters that minimize the divergence of predicted and measured expression intensities. They all consider the topology of the regulatory TF – gene network to be given and try to infer the actual TF activity present in a certain disease or under a specific experimental condition. Their primary output is a ranked list of TFs, sorted by their activity in a given group of samples. Several studies reported that such methods can be used to identify biomarkers for specific phenotypes in human cell lines and in vivo samples, for example in innate immunity, ageing related changes [Balwierz et al., 2014] or acute myeloid leukemia [Li et al., 2014].

Although certain evaluation steps were carried out for all methods, results in the original papers are not comparable as they used different input data sets, different background regulatory networks, and different evaluation metrics. Here, we implement a quantitative comparison including all of the previously presented methods to objectively analyze the results of different methods for estimating regulatory activity. We provide publicly available experimental data and regulatory networks as input to the methods to ensure transparency of our results.

Since the genome is regulated at multiple levels (see Chapter 2), the combination of different biological layers of information might help unraveling associations between biological entities and build elaborate markers of disease and physiology [Hasin et al., 2017; Huang et al., 2017]. Therefore, we use multi-omics patient data from three different cancer types as input to the methods (see Section 4.3). Furthermore, we evaluate the results of different methods by using mRNA expression data from knockdown and knockout experiments in human and E.coli cell lines (see Section 4.4). Altering the expression of single transcription factors offers an important source of information for estimating regulatory activity [Markowetz and Spang, 2007]. Additionally, we apply different underlying regulatory networks to investigate their influence on the results.

4.1. Evaluated Methods and Configurations

For the multi-omics data, which we downloaded from TCGA [Weinstein et al., 2013], we conduct the quantitative comparison for the method proposed by [Schacht et al., 2014],

RACER, RABIT and biRte. ISMARA is not included here, since it can only be used with its own, proprietary underlying regulatory network, and requires the upload of raw data which is prohibited by TCGA's terms of use. Also ARACNE is not included in the quantitative evaluation since it does not use background knowledge and we therefore consider its results as incomparable to the other methods.
In the case of knockdown data, which we retrieved from GEO [Edgar et al., 2002], we apply RACER, RABIT and biRte. Additionally, we run ISMARA for those data sets whose underlying experimental technologies are supported by the web service, even though the regulatory network is different. We use ARACNE to provide regulatory networks as input to the other methods, to study the influence of the underlying regulatory network. For the knockdown data sets, we do not apply the method by [Schacht et al., 2014], since its results from multi-omics are particularly poor.

The evaluated methods were described in detail in Chapter 3. Here, we briefly name the methods and their configuration considered for our quantitative comparison:

- RABIT published a C++ implementation which they provide on their website[1] and which we use with the FDR option set to 1. As RABIT takes differential expression into account, we use the difference of expression values between case and control group as input and order the TFs by t-value as proposed in the RABIT paper.

- For RACER we use the available R scripts[2] and extract the resulting sample-specific regulatory activities. To run RACER in the knockdown scenario, where only mRNA expression data is available, we set miRNA expression data, copy number variation and methylation scores, which have to be provided, to zero. The obligatory miRNA – gene network is artificially created where all dummy miRNAs and genes were connected. We compute separate models for case and control group and extracted the resulting sample-specific regulatory activities. TFs are ranked by their activity difference between the two groups.

- BiRte is available as a bioconductor R package. We use R version 3.3.2 with biRte version 1.10.0 and apply the method `birteLimma` to estimate regulatory activities with the options `niter` and `nburnin` set to 10000. As biRte has a randomized component, the resulting TF activities are not exactly the same for different runs. We average the final activity scores over 100 iterations of `birteLimma`.

- For the approach by [Schacht et al., 2014] we re-implemented their method as closely as possible to the original design using Python and the Cuneiform workflow language [Brandt et al., 2015; Bux et al., 2015]. Due to the high number of integer parameters in the original method, the complexity of optimizing the whole network at once would have by far exceeded our available computational resources.

[1] http://rabit.dfci.harvard.edu, accessed 10 September 2019
[2] http://www.cs.utoronto.ca/~yueli/racer.html, accessed 10 September 2019

Therefore, like in the original paper, we compute the model for each gene separately and restrict the number of regulating TFs per gene to six. As in [Schacht et al., 2014], we use the Gurobi Optimizer version 6.04, which is available under a free academic license. We compute separate models for case and control group and rank the TFs by their activity difference between the two groups.

- ISMARA is available via a web service[3]. We uploaded raw CEL files and grouped the samples according to their origin or treatment to compare the average regulatory activity between different conditions.

- For ARACNE, we use the implementation of the "minet" bioconductor package [Meyer et al., 2008] in R (version 3.38) and build the mutual information matrix with Spearman's correlation. The threshold for removing an edge in the `aracne` function was set to 0.1. We used ARACNE only to generate gene regulatory networks based on expression data and subsequently computed activity scores using any other of the presented methods.

4.2. Ranking

We compare the results of each method and in each data set by ranking the absolute values of the computed TF activity scores. The highest absolute activity value corresponds to rank 1. We appoint TFs that compared equal the same rank. Subsequently, a gap is left in the ranking numbers whose size is equal to the number of items that compared equal minus 1. Activities equal to zero are not considered. Therefore, the total number of ranked TFs is different in each method and data set.

We predominantly assess the rank of the TF that was knocked down. Additionally to the KD TF, we evaluate the ranks (if existing) of

- directly connected TFs in the network

- aliases provided in the GeneCards database version 4.8.0 Build 5[4] [Stelzer et al., 2016] for human TFs respectively synonyms from the EcoCyc database[5] [Keseler et al., 2017] for E. coli.

- TFs directly connected in a pathway from SignaLink 2.0[6] [Fazekas et al., 2013] for human TFs respectively from the EcoCyc database[5] [Keseler et al., 2017] for E. coli.

- TFs directly interacting with the KD TF according to TcoF-DB version 2.2.2[7] [Schmeier et al., 2017] for human.

[3] https://ismara.unibas.ch, accessed 10 September 2019
[4] www.genecards.org, accessed 10 September 2019
[5] www.ecocyc.org, accessed 10 September 2019
[6] www.signalink.org, accessed 10 September 2019
[7] http://tools.sschmeier.com/tcof/home, accessed 10 September 2019

For a given TF, we call all TFs in the union of these sets "related TFs". An overview is available in Table A.1 in the Appendix. The table shows all related TFs, irrespective whether they appear in our regulatory networks or not. For each method and data set individually, we evaluate whether the resulting ranks of all related TFs are significantly smaller than the average rank. We apply a one-sided one-sample t-test to compare the mean rank against the average rank (total number of ranked TFs divided by 2) and consider p-values < 0.05 as significant. Since the total number of t-tests is quite small (54) and nearly all p-values are above the significance level anyway, we do not apply multiple testing correction.

4.3. Validation using Multi-omics Data

In the original papers, the data sets used for evaluation vary between all methods. Therefore, we implement an evaluation framework to compare the method by [Schacht et al., 2014], RACER, RABIT and biRte in an objective and quantitative way. We use experimental data of three independent and publicly available data sets from TCGA [Weinstein et al., 2013] and a regulatory network as background knowledge.

4.3.1. Data Sets

We use experimental data from TCGA [Weinstein et al., 2013] for three cancer types: Colon adenocarcinoma (COAD), liver hepatocellular carcinoma (LIHC) and pancreatic adenocarcinoma (PAAD). For all three cancer types, mRNA expression, CNV, DNA methylation and miRNA expression data is available for primary tumor and normal tissue samples. These data sets are openly accessible via the NCI Genomic Data Commons Data Portal[8] or the NCI Genomic Data Commons Legacy Archive[9] (DNA methylation data) under the project names TCGA-COAD, TCGA-LIHC and TCGA-PAAD.

For mRNA gene expression we use processed RNA-Seq data in the form of FPKM (fragment per kilobase of exon per million mapped reads) values. The files include Ensembl Gene IDs which are converted to HGNC symbols using the Ensembl [Yates et al., 2016] BioMart tool[10] to match the IDs of the TF – gene network. In two cases, where multiple Ensembl Gene IDs map to one HGNC symbol, we choose the gene with highest $\log_2$ fold change between case and control group. miRNA expression is given as RPM (reads per million miRNA mapped) measurements. Both mRNA and miRNA data are centered using a weighted mean such that the mean of the case group equals the negative mean of the control group, and normalized via a weighted standard deviation. CNV data is retrieved as masked copy number segment. The Y chromosome and probe sets with frequent germline copy-number variation has already been removed by the data providers. Chromosomal regions are mapped to genes using the R package biomaRt

[8] https://portal.gdc.cancer.gov, accessed 10 September 2019
[9] https://portal.gdc.cancer.gov/legacy-archive, accessed 10 September 2019
[10] http://www.ensembl.org/biomart/martview, release 87, accessed 10 September 2019

[Durinck et al., 2009]. If multiple records map to one gene, the median of the segment mean values is calculated. For DNA methylation data we use the beta-values of Illumina Human Methylation 450 arrays as methylation scores. Multiple scores for the same gene are averaged within a sample.

We restrict our analyses to samples for which all four input data types are available. When multiple measurements for one sample and data type are available, we use only the first one in alphabetical order of the file name. After this selection procedure, 165 samples remain for COAD, 404 for LIHC and 180 for PAAD. A list including sample and file information is available online (additional file 1 of [Trescher et al., 2017])[11].

Together with the experimental data, all evaluated methods are given the same regulatory network as input. We use a publicly available human TF – gene network [Thomas et al., 2015] based on a text-mining approach available via the FastForward DNA database[12] and complemented it with TF – gene interactions from the public TRANSFAC database, release 7.0[13] [Wingender et al., 1996]. The network was built by text mining the entire Medline and an additional manual curation step of the top-ranking sentences. It thereby combines the content of regulatory databases with more than 300 validated regulatory relationships. This network includes 2894 interactions between 429 TFs and 1218 genes. The network is provided as additional file in [Trescher et al., 2017][11] and includes an adjacency list of the connected nodes of the TF – gene network. The list consists of three columns ("TF", "gene", "edge") where each row indicates an association with the value of "edge" between a TF and a gene. Complexes of TFs are indicated with a separating "." between their components.

4.3.2. Results

To ensure the result's comparability, we first use only mRNA expression data as input to the four methods. In a second evaluation, we include also other omics data sets where possible. We obtain lists with the regulators ranked according to the absolute value of their computed activity for each cancer type and method, with and without the use of additional inputs. For each cancer type we calculate the size of the overlaps in the four different results using the top 10 and top 100 regulators. The results for the top 10 regulators are shown in Table 4.1 (only mRNA) and Table 4.2 (multiple omics data sets). To better distinguish regulators and genes from cell lines or other abbreviations, we set TFs, miRNAs and genes in italics.

[11] https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-017-0419-z, accessed 10 September 2019

[12] http://fastforward.sys-bio.net, accessed 10 September 2019

[13] http://www.gene-regulation.com/pub/databases.html, accessed 10 September 2019

Only mRNA as Input

When only mRNA is used as input, only one TF is found by the three methods RACER, RABIT and biRte in each data set, respectively: *PHOX2B* for COAD, *EPAS1* for LIHC and *ELF1* for PAAD (see Table 4.1). A literature search of these TFs and their targets reveals clear associations to the respective cancer type. The TF obtained commonly for COAD, *PHOX2B*, is related to *TLX2*, a gene which has been shown to play a role in the tumorigenesis of gastrointestinal stromal tumors [Naumov et al., 2013]. *EPAS1*, which is found in the LIHC top 10 TFs of three methods, is linked to *CXCL12*, which plays an important role in metastasis formation of hepatocellular carcinoma by promoting the migration of tumor cells [Liu et al., 2008; Rubie et al., 2006]. For PAAD, three methods rank TF *ELF1* high, which is related to 14 genes in our network, inter alia to *BRCA2* and *LYN*. Mutations in the *BRCA2* gene have been implicated in pancreatic cancer susceptibility [Couch et al., 2007; Greer and Whitcomb, 2007], whereas the knockdown of *LYN* reduced human pancreatic cancer cell proliferation, migration, and invasion [Je et al., 2014]. These results underline that the methods are able to find biologically relevant information about regulation processes in cancer. Several TFs in the top 10 are found by two of the four methods. For instance, RACER and RABIT have four common top 10 TFs (*CDX2*, *NRF1* and *MYC* next to *PHOX2B*) in the COAD data set. The top 10 TFs found by the method by [Schacht et al., 2014] do not overlap with any top 10 TFs of the other methods in any data set. The agreement of RACER, RABIT and biRte in the top 10 TFs is statistical significant as the probability of finding common TFs in three sets of ten randomly chosen ones out of 429 TFs is below 0.006. Additionally, the methods do identify different TFs for different data sets, indicating the importance of the actual cancer specific mRNA expression values and that results are not dictated by the background network.

The number of overlapping regulators in the top 100 between the four methods and the three different data sets are shown in Figure 4.1. For RABIT, only 76 TFs for COAD (resp. 67 for LIHC and 57 for PAAD) can be ranked since all other TFs have an activity value equal to zero. When looking at the overlap of three of the four methods, the number of overlapping TFs is still the highest for the triplet RACER, RABIT and biRte. In the LIHC data set, two TFs are found in the top 100 of all four methods (*E2F4* and *SOX10*). *E2F4* is a downstream target of *ZBTB7*, which is associated to the expression of cell cycle-associated genes in liver cancer cells [Yang et al., 2012]. Two target genes of *E2F4*, *CDK1* and *TP73* are also involved in liver cancer development [Bisteau et al., 2014] and proposed as prognostic marker of poor patient survival prognosis in hepatocellular carcinoma [Stiewe et al., 2004]. Further, epigenetic alterations of the *EDNRB* gene, a target of *SOX10*, might play an important role in the pathogenesis of hepatocellular carcinoma [Hsu et al., 2006]. Even if the result of four methods finding two common TFs is not statistically significant (p-value=0.36), their association to liver hepatocellular carcinoma shows that the methods find at least a few relevant TFs.

However, when comparing different data sets, the methods tend to rank the same TFs under the top 100 to a greater or lesser extent. For example, the overlap of all top 100

Data set	Schacht et al.	RACER	RABIT	biRte
COAD	*INSM1*	*HOXA5*	*MYC*	*AHR**
	NR0B1	*SP4*	*KLF5*	*NR1I3**
	SNAI1	*MECOM*	*CDX2*	*KLF5*
	FOXC1	*MLXIPL*	*NRF1*	*PRDM1*
	PHOX2A	*CDX2*	*PRDM1*	*CDX1*
	FOXA1	*NRF1*	*NFYA*	***PHOX2B***
	SREBF2	*MYC.MAX.ZBTB17*	*NFKB1*	*ESRRA*
	NR4A1	***PHOX2B***	***PHOX2B***	*HOXA5*
	SNAI2	*HOXA10*	*RARG*	*TCF7L2*
	ARNT.HIF1A	*MYC*	*PITX2*	*SOX2*
LIHC	*NFIL3*	*GBX2*	*HNF4A*	*PHOX2A*
	NR0B1	*STAT5B*	*MYC*	***EPAS1***
	ELF2	*POU3F1*	*NRF1*	*HNF4A*
	NR4A2	***EPAS1***	*HNF1A*	*FLI1*
	ZNF384	*POU5F1*	*SP1*	*MTF1*
	INSM1	*ELK3*	*RARB*	*IKZF1*
	ATOH1	*PHOX2A*	*MTF1*	*NFATC1*
	SP4	*FOXF2*	*SOX10*	*POU3F1*
	KLF11	*MMP3*	*NR1I3*	*POU3F2*
	POU4F1	*GCM1*	***EPAS1***	*NFKB1*
PAAD	*RARB*	***ELF1***	*SPI1*	*SPI1*
	RBPJ	*SATB1*	*GATA2*	*PRDM1*
	USF1.USF2	*IRF1*	*PES1*	*PES1*
	BARX2	*STAT1.STAT2.IRF9*	*FOXO3*	*BACH1*
	USF2	*IKZF1*	***ELF1***	***ELF1***
	STAT3.STAT1	*NFATC2*	*RELA.REL*	*PURA*
	ETV4	*MYF5*	*NFE2*	*TFAP2B*
	HOXA1	*GATA2*	*CTCF*	*SATB1*
	STAT4	*NFYC*	*ATF1*	*NR2C2*
	ESR1	*PHOX2A*	*PURA*	*STAT1*

Table 4.1.: HGNC Symbols of the top 10 regulators found by each method for COAD (using 165 samples), LIHC (404 samples) and PAAD (180 samples) and the use of only mRNA data as input. TFs with equal activity values are marked with *. TFs found by several method's top 10 are marked in bold (when found by RACER, RABIT and biRte), blue (RACER and RABIT), red (RABIT and biRte) or yellow (RACER and biRte).

TFs of the three cancer types is only one TF for RABIT and nine TFs for biRte, but 16 TFs for the method by [Schacht et al., 2014] and even 32 TFs for RACER. Therefore, the results from RABIT and biRte seem to be more cancer type specific than the results from RACER and the method by [Schacht et al., 2014]. However, we do not specifically investigate the influence of the underlying network and its topology on the results here.

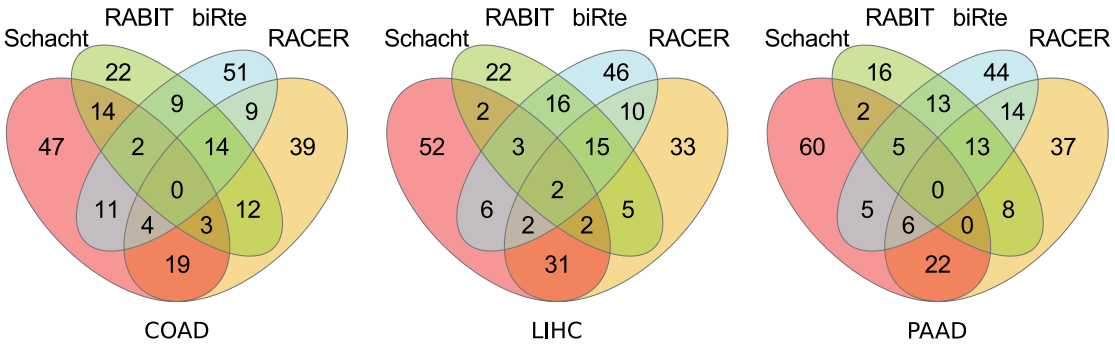


Figure 4.1.: Number of overlapping TFs in the top 100 of ranked TFs per method (for RABIT the overlap with the top 76/ 67/ 57 TFs (having activity > 0) in COAD/ LIHC/ PAAD is shown).

Multi-omics Data as Input

When not only taking mRNA into account but also miRNA, CNV and DNA methylation, the results are more difficult to compare, since every method uses a different way of combining different types of data. We are aware of the lower level of comparability of this approach regarding the multi-omics results in contrast to a scenario, where all methods are evaluated on the same set of input data. However, we here intend to use the maximum set of input data for each method to cover the effect of the use of multiple omics data sets compared to only mRNA as input.

BiRte is evaluated on mRNA and CNV data, RABIT on mRNA, CNV and DNA methylation data, and RACER additionally uses miRNA expression as input. Whereas RACER and RABIT consider CNV or DNA methylation data as one background factor and compute only one activity value, biRte evaluates the influence of each CNV separately.

The results (see Table 4.2) show that RACER exclusively ranks miRNAs high; not a single TF is found among the top 10 regulators. The influence of CNVs is high in LIHC and PAAD. The TFs that RACER found in the top 10 when using only mRNA data as input are still ranked high in the multi-omics scenario, e.g. the COAD top three TFs of the mRNA results are ranked 13th, 16th and 14th in the results of the multi-omics input. The difference of the results coming from the two input types is smaller for RABIT: Seven TFs are still in the top 10 for COAD (8 for LIHC and 6 for PAAD) when using CNV and DNA methylation additionally to mRNA data. Therefore, the contribution of additional input data seems not to be crucial for the performance of RABIT. BiRte considers each CNV as a potential regulator which increases the total number of regulators enormously. Still, two commonly present TFs in the top 10 of the COAD data set (even six for LIHC and one for PAAD) are found by either the sole mRNA input and the multi-omics approach. The overlap of the top 10 of RABIT and biRte in the multi omics case is significant with three TFs in LIHC (*HNF4A*, *EGR1* and

Data set	RACER	RABIT	biRte
COAD	*MIR130A*	*MYC*	*GUCA2A*
	MIR598	*NRF1*	*SLC25A34*
	MIR640	*KLF5*	*PLCD1*
	MIR554	*RARG*	*AHR*
	MIR921	*GFI1B*	*FAM163B*
	MIR631	*E2F1*	*NR1I3*
	MIR1202	*CDX2*	*KLF4*
	MIR548G	*NFYA*	*TRPM6*
	MIR602	*HOXA5*	*ADAMDEC1*
	MIR623	*PITX2*	*TMIGD1*
LIHC	*MIR187*	*HNF4A*	*PHOX2A*
	MIR892A	*EGR1*	*EPAS1*
	MIR638	*SP1*	*HNF4A*
	MIR517A	*NRF1*	*ADRA1A*
	MIR493	DNA methylation	*MTF1*
	MIR572	*MYC*	*IKZF1*
	CNV	*SOX10*	*EGR1*
	MIR192	*MTF1*	*FLI1*
	MIR1281	*RARB*	*CEBPB*
	MIR1244	*NR1I3*	*FOS*
PAAD	*MIR653*	DNA methylation	*RNU6-830P*
	MIR552	*SPI1*	*RN7SKP94*
	MIR381	*PES1*	*RNA5SP60*
	MIR668	*NFKB1.REL*	*SPI1*
	MIR587	*PURA*	*PHBP14*
	CNV	*NFE2*	*TOMM22P6*
	MIR596	*ATF1*	*IL22*
	MIR1180	*FOXO3*	*EEF1A1P24*
	MIR190B	*NFATC2*	*LINC01375*
	MIR216A	*IRF1*	*EIF4EP4*

Table 4.2.: HGNC Symbols of the top 10 regulators found by each method for COAD (using 165 samples), LIHC (404 samples) and PAAD (180 samples) and the use of multiple input data sets (RACER: mRNA, miRNA, CNV and DNA methylation; RABIT: mRNA, CNV and DNA methylation; biRte: mRNA and CNV). TFs found by several method's top 10 are marked in red (RABIT and biRte).

MTF1; p-value=0.001), but not significant with one TF in PAAD (*SPI1*; p-value=0.21). Three of them (*HNF4A*, *MTF1* and *SPI1*) were already found when using only mRNA data as input. Overall, the results for different input data sets show that the top ranked regulators are drastically changed when using additionally miRNA data in RACER, but change less when only CNV or DNA methylation data is provided in RABIT and biRte.

4.4. Validation using Knockdown Data

We suspect that the complexity of gene regulation in cancer is one reason for the questionable performance and low consistency of different methods' results in the previous analyses using multi-omics data. We therefore perform an additional experiment, where we focus on much less complex data and use knockdown experiments to evaluate different methods on estimating TF activity changes. We suppose that the highest change in activity will occur in the knocked down TF when comparing case and control samples. Many data sets of such high-throughput experiments for certain experimental conditions and different species have been published and are available in public repositories like GEO [Edgar et al., 2002] (Gene Expression Omnibus). In this straightforward and comparably simple setting, we expect that the methods are consistently able to identify the knocked down TF.

Here, we compare four different methods, namely biRte [Fröhlich, 2015], ISMARA [Balwierz et al., 2014], RABIT [Jiang et al., 2015] and RACER [Li et al., 2014], to infer transcription factor activity from gene expression data in knockdown (KD) experiments. We downloaded transcriptome data of four publicly available KD experiments from the GEO [Edgar et al., 2002] repository including different TF knockdowns in human and E.coli cell lines. To better distinguish KD TFs from cell lines or other abbreviations, we set TFs in italics.

4.4.1. Data Sets

We downloaded publicly available transcriptome data for different TF knockdowns in human and E. coli cell lines from the GEO repository [Edgar et al., 2002]. We chose three different experiments for TF silencing in human cell lines and one experiment in E. coli. The three experiments from human cell lines (GEO identifier GSE45838 [Alvarez et al., 2016], GSE17172 [Alvarez et al., 2009; Lefebvre et al., 2010] and GSE19114 [Carro et al., 2010]) contain data from 8 knocked down genes (*BCL6, FOXM1, MYB, bHLH-B2, FOSL2, RUNX1, C/EBPβ, STAT3*) and the double knockdown *C/EBPβ* & *STAT3*. The selected experiment in E. coli (GEO identifier GSE1121 [Covert et al., 2004]) comprises 5 knocked down genes (*AppY, ArcA, Fnr, OxyR, SoxS*) and the double knockdown *ArcA* & *Fnr*. Some of these experiments were conducted in several cell lines or conditions (see Section 4.4.1). Overall, we study 25 data sets (combinations of the experiment, the particular TF knockdown and different cell lines or growth conditions), 13 from human and 12 from E.coli. Throughout the section, we refer to the whole KD experiments from GEO as "experiments", which contain different KDs in cell lines or growth conditions, called "data sets". For an overview of the composition of the experiments, see Figures 4.3, 4.4 and Table 4.4.

PCA plots for all data sets are provided in Figure 4.2 showing the separation of treated and control samples. We map the given probe identifiers to HGNC Symbols (human data) or gene symbols from UniGene (E. coli). When multiple probes map to

one gene we compute a t-test comparing case and control group and keep the probe with smallest p-value.

GSE45838 [Alvarez et al., 2016] contains data from the knock-down of *BCL6* expression in human diffuse Large B-Cell Lymphoma cell lines. This experiment was performed in OCI-Ly7 and Pfeiffer GCB-DLBCL cell lines as triplicates, providing three case and three control samples per cell line. Gene expression was profiled on H-GU133plus2 Affymetrix gene chips. We analyze the samples in dependence of their cell line origin and treat them as two independent data sets since they are clearly separated in a PCA plot (see Figure 4.2, panel a).

GSE17172 [Alvarez et al., 2009; Lefebvre et al., 2010] consists of samples of Human Burkitt's lymphoma ST486 cells which were transduced either with non-target control shRNA lentiviral vectors, *FOXM1* shRNA or *MYB* shRNA lentiviral vectors (three samples in each condition). cRNA was hybridized in Affymetrix Human Genome U95 Version 2 Arrays. We use the MAS5 [Hubbell et al., 2002] normalized data as provided on GEO.

GSE19114 [Carro et al., 2010] includes 74 samples from knockdown experiments in human glioma cell line SNB19 and glioblastoma multiforme-derived brain tumor initiating cells (BTICs). shRNA-mediated silencing targeted *bHLH-B2*, *FOSL2*, *RUNX1*, *C/EBPβ* and *STAT3*. For SNB19, 10 control samples are available together with 4 samples with *bHLH-B2* knockdown, 4 with *FOSL2* knockdown and 3 samples each for *C/EBPβ*, *STAT3* and the combined *C/EBPβ* & *STAT3* knockdown. Data is available for *C/EBPβ*, *STAT3*, combined *C/EBPβ* & *STAT3* knockdown and a control condition for 11 samples in each group in BTICs. RNA was hybridized on Illumina HumanHT-12v3 expression BeadChip. Since the samples are clearly separated in a PCA plot by their cell type (see Figure 4.2, panel c), we treat data from SNB19 and BTICs independently.

GSE1121 [Covert et al., 2004] contains three samples of six E. coli strains with knockouts of transcriptional regulators in the oxygen response (*AppY*, *ArcA*, *Fnr*, *OxyR*, *SoxS* and the double knockout *ArcA* & *Fnr*) in both aerobic and anaerobic conditions. Additionally, three (aerobic condition) and four (anaerobic condition) wild type samples are available. Gene expression was profiled on Affymetrix E. coli Antisense Genome Arrays. We analyze the data from the two oxygen conditions independently.

As background network, we use two gene regulatory networks (one for human, one for E.coli) as input to the methods biRte, RABIT and RACER. Recall that ISMARA employs an own, inaccessible underlying network. The network including information on human regulatory relationships is based on a text-mining approach [Thomas et al., 2015] complemented with TF – gene interactions from the public TRANSFAC database,

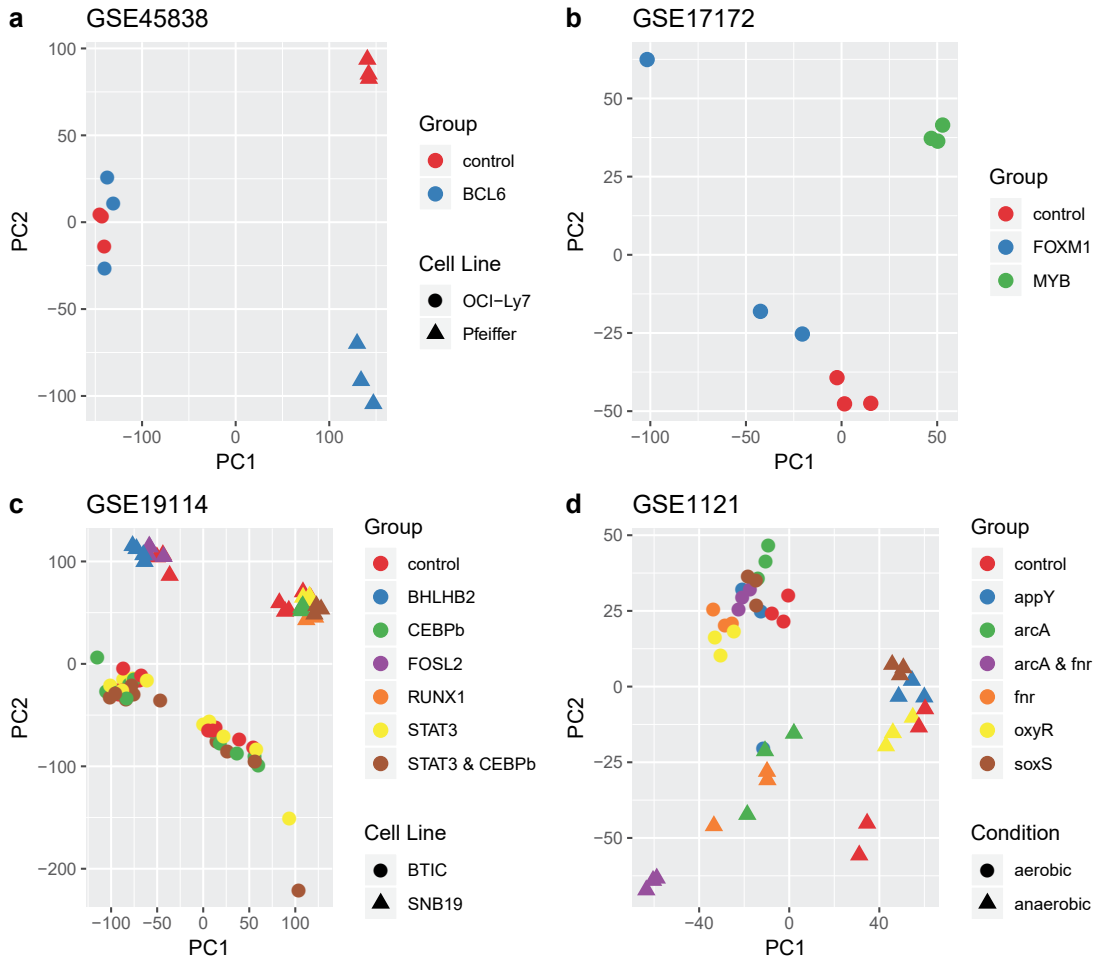


Figure 4.2.: PCA plots (showing first and second component) for all considered data sets. a) GSE45838 (*BCL6* knockdown), b) GSE17172 (*FOXM1* and *MYB* knockdown), c) GSE19114 (*C/EBPβ*, *STAT3*, *bHLH-B2*, *FOSL2* and *RUNX1* knockdown), d) GSE1121 (*ArcA*, *AppY*, *Fnr*, *OxyR* and *SoxS* knockout)

release 7.0[14] [Wingender et al., 1996] (see Section 4.3.1 for a more detailed description). The network for E. coli was retrieved from RegulonDB, version 9.0, Release 9.4 [Gama-Castro et al., 2016]. We keep those interactions for which at least one entry in the column "Evidence that supports the existence of the regulatory interaction" is mentioned. The network contains 4273 interactions between 206 TFs and 1798 genes.

[14]http://www.gene-regulation.com/pub/databases.html, accessed 10 September 2019

4.4.2. Results

We predominantly assess the rank of the TF that was knocked down and the total number of ranked TFs. We additionally check for aliases and determine the ranks of neighbor TFs in the network, of co-members in a pathway and of interacting TFs. To examine whether the methods are able to detect a common signal in the data, we compare the overlap of the top 100 ranked regulators of all methods within one data set. We additionally perform activity estimation on smaller networks and on networks inferred de-novo by ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [Margolin et al., 2006] to assess the influence of the network on the results. To test whether the mere differential expression is a better predictor for TF activity changes, we calculate ranks of differential expression for the knocked down TFs and compare those to the activity ranks.

Our results show that, although almost all KD TFs show differential expression, their activity ranks are only in 15 out of 54 cases within the top 5% of all ranked TFs (compare Table 4.4). In E. coli, the identification of the KD TF by activity estimation yields slightly better results compared to human cell lines. When looking not only at the KD TF but also at regulators related to the KD TF in the network or a pathway, we identify only a single case where the mean of the ranks of all related TFs is significantly smaller than expected by chance. The overlap of the top 100 ranked regulators of all methods within one data set is small and statistically insignificant. The reduction of the network size or the use of ARACNE's inferred networks does not improve the results.

Differential Expression

First, we test whether the knocked down TFs themselves are differentially expressed, which is the case for all human KD TFs except *C/EBPβ* in BTICs (brain tumor initiating cells), both in the single KD and double KD together with STAT3, see Figure 4.3. Unexpectedly, the expression of RUNX1 is significantly upregulated in SNB19 case samples compared to the control samples. Nonetheless, we include RUNX1 in our analyses since we are only interested in finding absolute TF activity changes. In E. coli (see Fig. 4.4), all KD TFs are significantly downregulated in the corresponding case samples. The according p-values are given in the Appendix A.2.

Additionally, we check whether the differential expression per se would be a good predictor for the determination of knocked down TFs in a data set. Therefore, we compute differential expression separately in each data set, contrasting the expression of the corresponding case and control samples and evaluated the ranks of differential expression for the KD TFs via a two-sided t-test. We rank the genes according to the p-value of the t-test (smallest p-value corresponds to rank 1). We do not apply any multiple test correction, since we are not interested in the precise p-value but only the order of p-values to assign ranks. The results are shown in Fig. 4.3. In human, in 9 out of the 13 data sets, the TFs are ranked within the top 5%. In E. coli, the number of KD

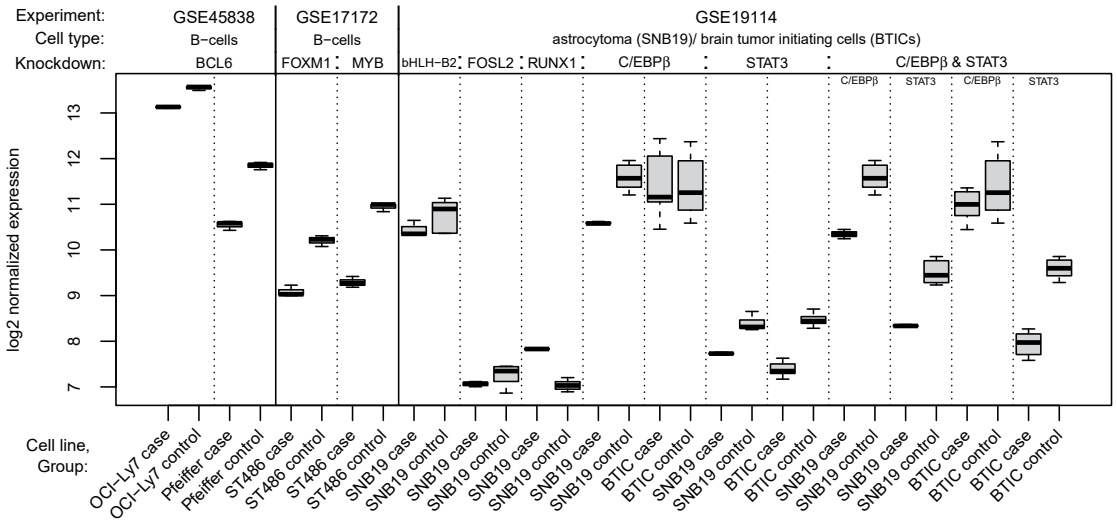


Figure 4.3.: Boxplots of $\log_2$ normalized expression values for all human KD TFs, comparing respective case and control groups. For the double KD *C/EBPβ* & *STAT3*, separate boxplots for each TF are shown. In all experiments, expression in case samples is significantly lower than in control samples, except for *C/EBPβ* (single and double KD) in BTICs and *RUNX1* KD.

Figure 4.4.: Boxplots of $\log_2$ normalized expression values for all E. coli KD TFs, comparing respective case and control groups. For the double KD *ArcA* & *Fnr*, separate boxplots for each TF are shown.

TFs in the top 5% TFs is 3 in the aerobic and 5 in the anaerobic condition out of 6 data sets in either condition. As expected, the KD TFs are in about two third of the considered data sets amongst the TFs with the highest changes in differential expression.

Organism	Experiment	TF knockdown	Cell line/ condition	rank	total
Human	GSE45838	***BCL6***	OCI-Ly7	9	371
			Pfeiffer	5	371
	GSE17172	***FOXM1***	ST486	1	331
		MYB	ST486	1	331
	GSE19114	***bHLH-B2***	SNB19	174	368
		FOSL2	SNB19	108	368
		RUNX1	SNB19	72	368
		C/EBPβ	SNB19	7	368
			BTICs	330	368
		STAT3	SNB19	1	368
			BTICs	1	368
		C/EBPβ & STAT3	SNB19	6 / 4	368
			BTICs	117 / 1	368
E. coli	GSE1121	***AppY***	aerobic	8	150
			anaerobic	4	150
		ArcA	aerobic	13	149
			anaerobic	7	150
		ArcA & Fnr	aerobic	11 / 4	151
			anaerobic	13 / 8	151
		Fnr	aerobic	3	151
			anaerobic	3	150
		OxyR	aerobic	1	149
			anaerobic	1	150
		SoxS	aerobic	17	150
			anaerobic	1	150

Table 4.3.: Ranks for differential expression of KD TFs and total number of ranked TFs per data set. Differential expression ranks of KD TFs in the top 5% of all ranked TFs are marked in dark orange, ranks in the top 5-10% in yellow and ranks in the top 10-20% in light orange. Two ranks in one table cell refer to a combined KD of two TFs and are given in the order of the TFs at the beginning of the table row.

Ranking of knocked down TFs

We next apply biRte, ISMARA, RABIT and RACER to determine the respective KD TFs' ranks. Since neither Illumina chips nor the Affymetrix E. coli Antisense Genome Array are supported, we can run ISMARA only on the data sets with *BCL6* (GSE45838) and *FOXM1*/*MYB* (GSE17172) knockdown. The KD TFs are only in 15 out of 54 cases within the top 5% of all ranked TFs (4 out of 18 in human and 11 out of 36 in E. coli). Of the 54 cases where non-zero ranks are computed, 27 result from biRte, one from

ISMARA, 13 from RABIT and 13 from RACER. Due to stringent filtering thresholds within the methods, no activity score is assigned to the KD TF in 37 cases. The resulting ranks of knocked down TFs and the total number of ranked TFs per method and data set are shown in Table 4.4. Favorable results, meaning that the knocked TF is highly ranked, are marked in green.

We observe that biRte assigns at least some activity for nearly all KD TFs. In 3 out of the 13 human data sets where ranks are specified, biRte ranks the knocked down TF in the top 5% (*FOXM1*, *RUNX1* and *STAT3* in SNB19). In E. coli, the results from biRte are better with 8 out of 14 TFs in the top 5% and another two TFs in the top 10%. In all other data sets, the ranks for the TFs in question are quite low. ISMARA could only be applied to GSE45838 and GSE17172 since the chips from the other experiments are not supported by the online interface. In one data set (*MYB*) the KD TF is highly ranked (10th out of 602), but ISMARA does not provide any ranks for the two other KD TFs (*BCL6* and *FOXM1*). Since the underlying network from ISMARA is not accessible, we cannot discern whether the TF is not present in the network or is not considered important by the ranking procedure. RABIT removes TFs with insignificant cross-sample correlation from the results and therefore only provides the ranks of, on average, 56 TFs in our analyses. It does not provide any activity score for the KD TF in over half of the data sets (12 in human, 4 in E. coli). In human, not a single KD TF is ranked in the top 20%. However, in E. coli, RABIT is able to identify *AppY* (rank 1) and *ArcA* as knocked down TFs in the anaerobic condition (rank 2 in the single KD and rank 1 in the combined KD *ArcA* & *Fnr*). In contrast, RACER ranks only one KD TF for the human data sets at all (*BCL6*) and does not rank any KD TF highly in E. coli. In some human data sets, RACER even reports zero total active regulators.

Related TFs

We expect that the knockdown of a certain TF does not only affect the activity of this TF itself, but also influences the activity of related TFs. Therefore, for each KD TF, we determine the ranks of a set of related regulators. We define as related all TFs directly connected in the same pathway (information from SignaLink [Fazekas et al., 2013] for human respectively EcoCyc [Keseler et al., 2017] for E. coli), direct neighbors in the TF – gene network, directly interacting TFs (information from TcoF-DB v2 [Schmeier et al., 2017], human) and presumed aliases from the GeneCards [Stelzer et al., 2016] (human) and EcoCyc [Keseler et al., 2017] database (E. coli). An overview of the related TFs can be found in A.1 in the Appendix.

We show the resulting ranks and according p-values of the KD TF and related TFs for one exemplary result (*MYB* KD from GSE17172) in Table 4.5. All other results are given in Table A.3 in the Appendix. We observe that related TFs are rarely ranked highly by any of the methods. Only one related TF (*JUN*), which is directly connected to *MYB* in the human regulatory network, is ranked among the top 20% TFs by two of the four methods (biRte rank 50, ISMARA rank 23). Previously, it was shown that

Organism	Experiment	TF knockdown	Cell line/ condition	biRte rank	biRte total	ISMARA rank	ISMARA total	RABIT rank	RABIT total	RACER rank	RACER total
Human	GSE45838	***BCL6***	OCI-Ly7	266	404	-	500	-	58	-	88
			Pfeiffer	163	405	-	500	-	53	68	143
	GSE17172	***FOXM1***	ST486	9	398	-	602	-	63	-	4
		MYB	ST486	112	404	10	602	19	46	-	0
	GSE19114	***bHLH-B2***	SNB19	186	402			-	42	-	0
		FOSL2	SNB19	355	404			-	54	-	0
		RUNX1	SNB19	8	401			37	49	-	0
		C/EBPβ	SNB19	-	404			-	49	-	0
			BTICs	328	397			-	61	-	14
		STAT3	SNB19	4	403			29	59	-	0
			BTICs	209	405			-	60	-	14
		C/EBPβ & STAT3	SNB19	-/ 31	400			-/-	51	-/-	0
			BTICs	402/188	410			-/-	71	-/-	14
E. coli	GSE1121	***AppY***	aerobic	119	199			-	48	73	152
			anaerobic	15	198			1	43	71	121
		ArcA	aerobic	198	198			-	32	70	142
			anaerobic	1	199			2	42	135	147
		ArcA & Fnr	aerobic	6 / 7	199			5/6	29	108/-	133
			anaerobic	1 /148	198			1 /-	45	34/104	115
		Fnr	aerobic	9	199			10	33	-	137
			anaerobic	192	199			43	55	127	143
		OxyR	aerobic	7	197			28	34	83	135
			anaerobic	6	199			10	35	94	121
		SoxS	aerobic	1	199			10	40	95	146
			anaerobic	14	199			-	45	92	119

Table 4.4.: Ranks of knocked down TFs and total number of ranked TFs per method and data set. Ranks in the top 5% of all ranked TFs are marked in green and ranks in the top 5–10% in light green. Two ranks in one table cell refer to a combined knockdown of two TFs and are given in the order of the TFs at the beginning of the table row. An empty table cell (in ISMARA column) indicates that the method was not applicable to the data set. A dash is shown when a TF was not ranked by a method (see text for explanation of different numbers of ranked genes).

JUN contributed to the transcriptional activation of *MYB* [Nicolaides et al., 1992; Vorbrueggen et al., 1994]. For each method and data set individually, we evaluate whether the mean of the resulting ranks of all related TFs is significantly smaller than the average rank expected at random (total number of ranked TFs divided by 2). Only one out of 54 of the mean ranks of the estimated activity changes is significantly below the average rank: In E. coli, biRte ranks *OxyR* and a related TF highly in the anaerobic condition ($p = 0.002$). However, since this result is obtained with quite a small sample set (only two ranked TFs), we consider it not representative.

TF	biRte	ISMARA	RABIT	RACER
MYB	**112**	**10**	**19**	**-**
ETS1	-	292	-	-
HOXA9	2	431	27	-
IRF1	386	-	-	-
JUN	50	23	10	-
JUND	70	234	-	-
GATA3	34	580	25	-
MYC	130	-	-	-
NR3C1	129	299	24	-
PAX5	353	232	-	-
PAX6	297	411	-	-
SNAI2	122	-	-	-
SP3	270	196	-	-
HLF	203	-	-	-
MAF	56	-	-	-
SMARCA2	314	-	-	-
SP100	-	209	-	-
total	404	602	46	0
p-value	0.160	0.284	0.274	-

Table 4.5.: For experiment GSE17172: Ranks of *MYB* (bold) and related TFs, total number of ranked TFs per method and p-value indicating significance of test whether the mean of the ranks of all related TFs is smaller than the average rank. Ranks of TFs in the top 5% of all ranked TFs are marked in dark green, ranks in the top 5-10% in green and ranks in the top 10-20% in light green. When a TF was not ranked, "-" is shown.

Overlap

Since the ranks of knocked down and related TFs are quite different in each method, we examine whether the methods might detect a common signal in the data such as a drastic change elsewhere in the network incurred by the KD. To this end, we compare the overlap of the top 100 ranked regulators of all methods within one data set.

We find very little overlap in human cell line data. The highest overlap among three methods (biRte, RABIT and ISMARA) occurs in *FOXM1* with only four common TFs within the top 100 (*JUN*, *MYBL2*, *NR2F2* and *FOXO4*). These results make sense, as the expression of *FOXM1* and *MYBL2* as its downstream factor are significantly associated with clinical stages and overall survival of glioma patients [Zhang et al., 2017] and is very high in Burkitt lymphoma [Höglund et al., 2011]. Further, *MYBL2* deregulation occurs in a broad spectrum of cancer entities [Musa et al., 2017; Sadasivam et al., 2012]. *FOXM1* is a direct target of repression by *FOXO* proteins. An inactivation of *FOXO* or overexpression of *FOXM1* is associated with tumorigenesis and cancer progression [Wilson et al., 2011]. Nevertheless, the overlap is small and not significantly larger than expected at random ($p = 0.81$, obtained by simulating the size of the overlap of three lists when sampling 100.000 times 100 out of 429 TFs per list).

In E. coli, the number of common TFs from biRte, RABIT and RACER is higher, but also not significantly ($p = 0.96$), with a maximum overlap of 18 TFs (*ArcA* & *Fnr* knockdown in the anaerobic condition). The overlap contains, for example, *ArcA*, which is activated in anaerobic conditions [Compan and Touati, 1994], *NtrC*, which is shown to be upregulated during the transition from anaerobic to aerobic conditions [Partridge et al., 2006], and *AdiY*, which is maximally induced under anaerobic conditions [Stim-Herndon et al., 1996]. Although the methods do not find the knocked down TF itself, at least in our E. coli data sets they commonly find TFs biologically relevant for the condition under consideration. The results are exemplarily shown for *FOXM1* and the combined *ArcA* & *Fnr* KD (anaerobic condition) in Figure 4.5 and in A.4 in the Appendix for all other TFs.

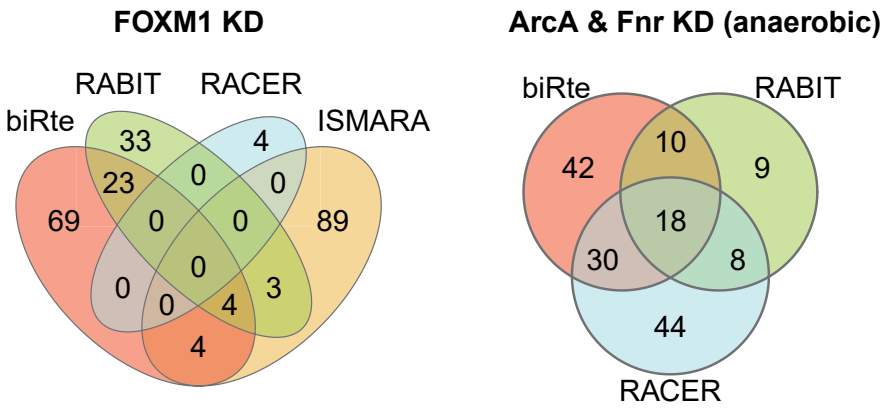


Figure 4.5.: Number of overlapping TFs in the top 100 by estimating TF activity with different methods. Venn diagrams are shown for *FOXM1* knockdown in human (left) and for the combined *ArcA* & *Fnr* knockdown in E. coli for the anaerobic condition (right). For RABIT and RACER, the total number of ranked TFs was below 100 in some cases (see Table 4.4).

Network Alterations

The previous results showed that, in a few cases, the methods were able to find biologically plausible information, although they did not identify the knocked down TF or its functional vicinity. One possible reason for this observation, which is in contrast to results published with the methods [Fröhlich, 2015; Jiang et al., 2015; Li et al., 2014], is that the regulatory networks used in the original work were much smaller compared to our networks. To assess whether the usage of a smaller network improves the results, we restrict the underlying TF – gene network to the neighborhood of each knocked down TF with a distance of two. Note that this design gives a very favorable prior to the analysis. An exemplary restricted network for *FOSL2* is presented in Figure 4.6. We apply biRte, RABIT and RACER again using these individual smaller networks for the human data sets and perform TF ranking. The resulting TF activities are shown in Table 4.6 and are not better than for the full networks. Only *RUNX1* and *STAT3* are ranked within the top 5% and *FOXM1* in the top 10% using biRte. This result was already obtained using the full network (compare Table 4.4). We conclude that the use of smaller and more focused regulatory networks alone is not sufficient to obtain more accurate results in human.

Organism	Experiment	**TF KD**	Cell line/ condition	biRte		RABIT		RACER	
				rank	total	rank	total	rank	total
Human	GSE45838	***BCL6***	OCI-Ly7	41	53	8	12	-	24
			Pfeiffer	31	53	-	5	-	0
	GSE17172	***FOXM1***	ST486	6	97	-	7	-	0
		MYB	ST486	41	156	12	19	-	21
	GSE19114	***bHLH-B2***	SNB19	20	63	-	7	-	5
		FOSL2	SNB19	16	26	-	1	9	26
		RUNX1	SNB19	1	43	-	4	1	3
		C/EBPβ	SNB19	-	95	-	14	-	0
			BTICs	93	95	-	17	-	3
		STAT3	SNB19	4	105	29	29	-	4
			BTICs	71	105	-	16	-	20

Table 4.6.: Ranks of KD TFs and total number of ranked TFs per method and data set for the restricted networks. Ranks of KD TFs in the top 5% of all ranked TFs are marked in green and ranks in the top 5–10% in light green. When a TF is not ranked, "-" is shown.

To further study the influence of the underlying regulatory network, we apply the popular method ARACNE [Margolin et al., 2006] to reconstruct ab initio a gene regulatory network from the given transcriptome data exemplarily for the *FOXM1* KD (human) and *AppY* KD (E. coli). We use the resulting gene regulatory networks as input to the investigated TF activity estimation methods and rank the resulting TF activity scores. Although the networks inferred by ARACNE have a higher density compared to our

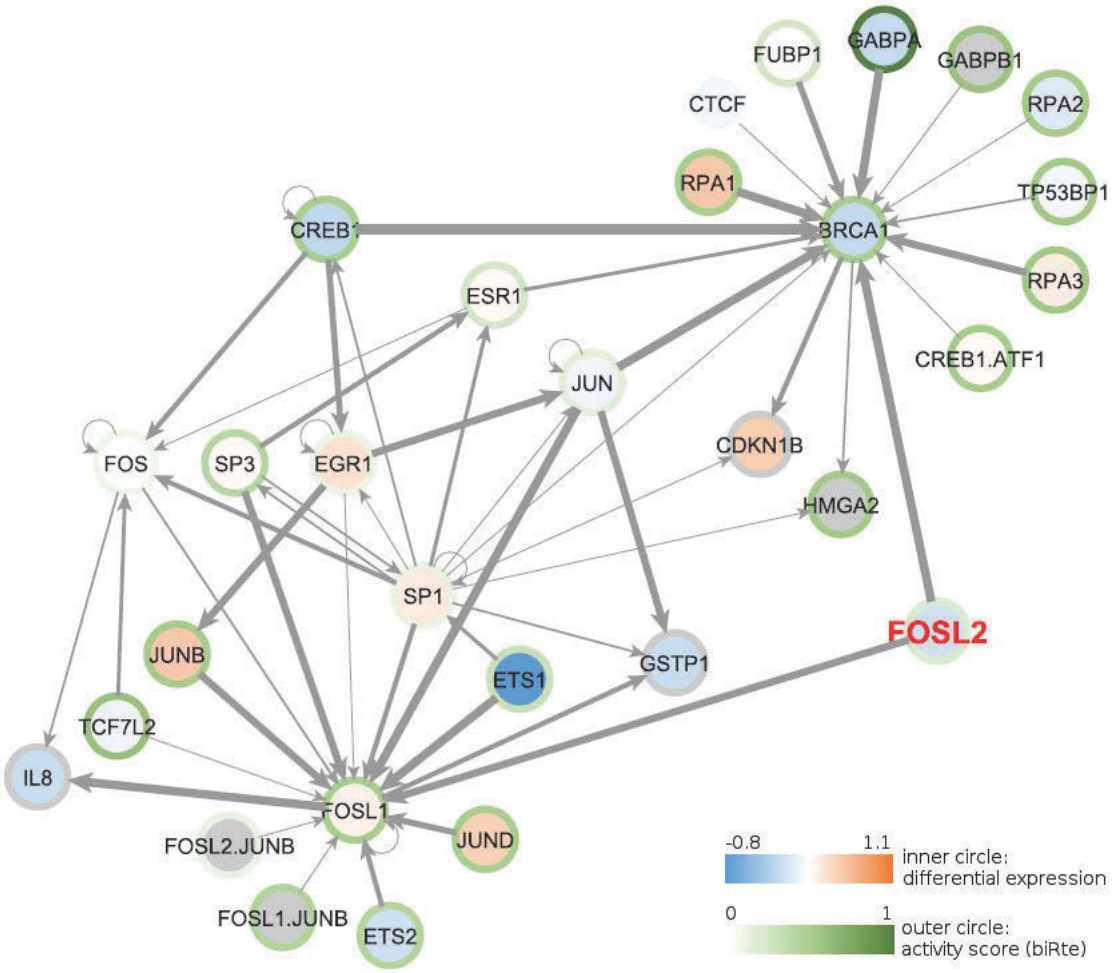


Figure 4.6.: Restricted network for *FOSL2*. The color of the inner circle corresponds to the differential expression of case vs control samples from GSE19114, SNB19 cell line with *FOSL2* knockdown ($\log_2$ fold changes): Blue colors correspond to downregulated, red colors to upregulated genes in the case samples; genes with missing expression are colored in gray. The color of the outer circle corresponds to the inferred activity score from biRte, ranging from 0 (no activity, white) to 1 (high activity, dark green). The edge width corresponds to the absolute correlation of the expression values between the two adjacent nodes: Small absolute correlation values are marked with a thin line, higher absolute correlation values with bolder lines. Edges with missing correlation values and self-correlation are given the thinnest line width.

original networks (see Appendix A.5), the resulting TF activity rankings are comparable (see Table 4.7). Therefore, the network provided as background knowledge to the methods seems not to be the most important element to explain the overall bad performance.

Organism	Experiment	**TF KD**	Cell line/ condition	biRte rank	biRte total	RABIT rank	RABIT total	RACER rank	RACER total
Human	GSE17172	***FOXM1***	ST486	9	248	-	70	-	178
E. coli	GSE1121	***AppY***	aerobic	15	145	-	19	21	103

Table 4.7.: Ranks of KD TFs (bold) and total number of ranked TFs per method using a network inferred by ARACNE as input. Ranks of TFs in the top 5% of all ranked TFs are marked in green and ranks in the top 5-10% in light green. When a TF was not ranked, "-" is shown.

4.5. Discussion

We conducted a comparative evaluation of different transcriptome-based TF activity estimation methods using multi-omics and knockdown data sets. Our results are easily reproducible since they are based on publically available data sets, networks and methods (except for the method by [Schacht et al., 2014], which we only used in our analyses of multi-omics data). The results show that the methods are able to find biologically relevant information about regulation processes in cancer. The overlap of results from different methods evaluating a specific data set is partly significant. The methods rank different regulators highly in different data sets, pointing to the importance of the actual cancer specific mRNA expression data and emphasizing that the results are not only dependent on the background network. However, the results of different methods vary greatly, which is also reflected by our results evaluating the knockdown data sets: We showed that estimates of TF activity are not quite robust since only in around a fourth of all cases the KD TF is ranked within the top 5%, despite that the KD TFs themselves are differentially expressed. In many cases, the methods did not assign any activity for the KD TF due to the internal filtering.

4.5.1. Networks

We used a gene regulatory network constructed by a text mining approach [Thomas et al., 2015] and complemented it with TF – gene interactions from the public TRANSFAC database [Wingender et al., 1996]. The construction of the text mining network included an extensive manual curation step, to improve the reliability of the detected relations compared to a completely automated approach. In addition to the text-mining, the network also contains interactions reported in the TRANSFAC database, which is based on biological experiments. For the E. coli KD data, the network was retrieved from RegulonDB [Gama-Castro et al., 2016], a gold standard in the field. We therefore believe that both the human and the E. coli network represent a pertinent choice to provide background knowledge to the methods. Further, for the KD data, the use of other networks (restricted versions of the original human and E. coli networks or networks

inferred by ARACNE) did not improve or change substantially the results. Also, the network size is not a negative factor for prediction performance, since the use of smaller networks did not improve the detection of KD TFs for any of the activity estimation methods. We conclude that the results are not imposed by the network given as input to the methods.

However, we did not evaluate the variations of topological structure of the networks, the effects of incompleteness, or different error rates. These aspects are known to have a severe influence in network analysis [Babtie et al., 2014; Luscombe et al., 2004] and it would be interesting to assess the behavior of the methods in these cases, for example via simulation studies. Hence, we later analyze the influence of randomized network edges in an artificial network on activity estimation (see Chapter 5). Further, larger TF-gene networks could be used as input for the methods, such as RegNetwork[15] [Liu et al., 2015], which contains 1456 human TFs, 1904 miRNAs and 19719 target genes from 25 databases. A network covering lung-specific TFs and their predicted targets, LungNet, was constructed by [Chen et al., 2017]. They were able to show that lung-specific TFs became consistently and preferentially inactivated in lung cancer, in precursor lung cancer lesions and partly in normal cells exposed to smoke carcinogens. This network could be useful in the analysis of lung-disease specific data sets, e.g. from TCGA. A recent study by [Garcia-Alonso et al., 2019] investigated in three data sets how background networks affect estimated TF activity using VIPER [Alvarez et al., 2016]. They conclude that literature curated information is the best source of information and assembled a collection of TF-target interactions for 1541 human TFs together with confidence scores.

4.5.2. Data Sets

In general, the selection of experiments affects the outcome of the methods. We use experiments from TCGA and the GEO platform, established and extensive repositories for omics data sets, to ensure an easy and public access to the data and to allow other researchers to replicate our results. We chose data from different species, different diseases, different cells of origin and cell lines, from various contributors and data measured on different arrays to make our results less dependent on a specific data sets. The chosen experiments had to fulfill certain criteria: For the evaluation based on omics data from cancer patients we only chose cases for which all required data types were available. For the knockdown data we chose to include only experiments with at least three samples per condition. Further, as we wanted to include ISMARA as a method for estimating TF activity, we preferred experiments whose Affymetrix chips were supported by its web service. These constraints limit the number of possible data sets and the use of other experimental data, different underlying networks or additional methods might produce different results. However, since we draw our conclusions from a total of three cancer and 25 KD data sets, amongst which we do not detect a pattern justifying an especially good or bad performance, we believe that our results show not only individual artefacts but are generalizable to the estimation of TF activity.

[15] available online via www.regnetworkweb.org, accessed 14 August 2019

We can also exclude the number of samples within a data set as a restricting element, as data sets with more samples do not achieve better results than those with fewer samples. For example, in some of the experiments we chose from GEO [Edgar et al., 2002], the sample size is relatively small with on average 4 case and 6 control samples per data set and a partly high variation within the groups (compare Figures 4.3 and 4.4). However, even in the data sets with larger sample size or with smaller variation, the method's results are not better compared to the less favorable data sets.

4.5.3. Performance across Methods

The comparability of different methods is only given when the same experimental data and a common regulatory network are provided as input. For the data based on cancer patient data, we therefore mainly compared the results based on mRNA expression data. Unfortunately, the multi-omics results are less comparable across different methods, since they all use a different combination of input data sets due to different model structures. However, we intended to use the maximum number of experimental input data sets to assess the influence of the use of additional omics data, allowing only the comparison of results from one method in this case. We observe that the results from RACER change greatly when integrating additional omics data, whereas the results from RABIT and biRte vary less. Further, the incorporation of many data types leads to an increasing number of parameters in the models, resulting in complex designs, which are prone to overfitting. Only ISMARA and biRte include explicit parameter priors to address this problem. Since the results from the method by [Schacht et al., 2014] have the least overlap with all other methods and no implementation is publically available, we chose to exclude this method from further investigation regarding the KD data sets.

When comparing the results of different methods by searching the literature for commonly found TFs, we inherently can only find already existing knowledge, restricting the explanatory power of our analyses. The closest comparable evaluation effort we are aware of addressing gene regulatory network reconstruction in the "DREAM5 – Network Inference" challenge [Marbach et al., 2012], but no generally accepted benchmarks are available to compare the results of methods estimating regulatory activity. [Berchtold et al., 2016] published a method called i-score to assess the target genes whose changes are strictly inconsistent with the predicted activity states of their corresponding TFs. They also found that active TF predictions were very different across methods, when comparing ISMARA [Balwierz et al., 2014], plsgenomics [Boulesteix and Strimmer, 2005], DREM [Ernst et al., 2007] and T-profiler [Boorsma et al., 2005]. They concluded that for many genes it is not possible to explain the observed effects with the current networks, likely because of missing edges in the network. We therefore focused on much less complex data, using knockdown experiments in human and E.coli to evaluate different methods on estimating TF activity changes. Supposing that the highest change in activity will occur in the knocked down TF we expect to partly circumvent the evaluation problem of unknown results.

4.5.4. Knockdown

The poor overall performance concerning the KD data sets cannot be attributed to a low effectiveness of the knockdown, which has an enormous effect on the TF's gene expression: Nearly all KD TFs show a significantly high differential expression and most of them have one of the highest changes in differential expression of all genes in the respective data set. We expect the methods to recognize such a drastic change in expression and activity represented by the KD. However, they can only rarely find the KD TF even when its differential expression is very high. This might indicate, that the KD itself affects only a small portion of the whole gene expression. Then one could argue, that the methods do not detect such particular changes and seem to be robust against limited variation in the input data. Nonetheless, the KD signal is clearly present in the data and expected to be found by the methods.

4.5.5. Human vs. E. coli

The results from E. coli are better compared to the results from human cell lines, both regarding the detection of the KD TF and regarding the agreement among different methods. The gene regulatory network of E. coli is probably the best characterized one of all species [Fang et al., 2017] with a gold standard of experimentally validated interactions from RegulonDB [Gama-Castro et al., 2016]. Even under such optimal conditions, the obtained results have only a poor quality. Conversely, a comprehensive characterization of the human regulatory repertoire is lacking since only about half of the estimated 1,500–2,000 TFs in the mammalian genome is known [Vaquerizas et al., 2009] and the existing knowledge about regulatory effects is scattered over the biological literature and different, partly commercial, databases, impeding the construction of comprehensive networks [Thomas et al., 2015]. We expected that the estimation of TF activity in human is a much harder task compared to its estimation in E. coli, which is partly confirmed by our results.

We also examined whether the methods were able to detect a common signal in the data at all and compared the overlap of the top 100 ranked regulators of all methods within one KD data set. The overlap in human data is quite small, but consistently larger in E. coli. We attribute the low similarity of the results partly to the noisy character of the transcriptome data provided as input. Also, many other factors important for regulation, like chromatin structure or post-transcriptional effects, are ignored. However, in both human and E. coli, the intersection of methods identifies some biologically plausible TFs for the condition under consideration. In the literature, we find many examples of such evaluation procedures [Balwierz et al., 2014; Fröhlich, 2015; Jiang et al., 2015; Li et al., 2014], where highly ranked TFs are found to be biologically important.

4.6. Conclusion

Except for a study from [Garcia-Alonso et al., 2019], analyzing the influence of the quality of TF–target interaction data sets on the estimation of TF activities, we are not aware of any other independent study on the performance of optimization-based algorithms for the estimation of whole genome transcription factor activity. Our results compare inter alia the performance on multi-omics data sets. We used a publicly available human TF – gene network [Thomas et al., 2015] together with experimental data from TCGA [Weinstein et al., 2013] for three cancer types to identify key biomarkers for these specific diseases. The results show that all methods seem to detect strong signals and find biologically relevant information about regulation processes in cancer, but sensitivity is low and the mutual result overlaps from different methods are small, though sometimes statistically significant. This seems surprising as all methods essentially follow the same goal, i.e., identification of the most differentially active TFs or genes. This low coherence in the results of different methods led us to the new experimental design of using knock-down. However, also on this presumable much simpler problem the result overlaps are very low and the knocked-down transcription factor is only very rarely identified. In the knockdown scenario, the investigated methods for estimating TF activity are not able to robustly detect knocked down TFs neither in human nor in E. coli data. We believe that the main reason for this deficiency is the simplistic model of cellular processes used even in the more complex methods like ISMARA. We can only speculate which aspects are primarily responsible for the limited performance. All considered methods only use gene expression data whereas other important regulatory processes such as epigenetic mechanisms like DNA methylation, chromatin remodeling, complex promoter structures, and post-transcriptional regulatory processes via microRNAs are disregarded. The inclusion of further data types would probably change the outcome of the methods and might improve results. Also, all models assume linear relationships between TFs and lack a notion of kinetic or temporal effects [Klinger and Blüthgen, 2018]. Although time series expression data from TF knockdown or TF induction experiments exist [Atger et al., 2015; Nishiyama et al., 2009], the selected methods cannot make use of this type of data. Another possible reason for the failure of the methods might be their inability to model TF self-regulation and feedback loops despite their known importance for gene regulation [Alon, 2007; Komili and Silver, 2008]. We therefore propose a new method for estimating transcriptional activity with a particular focus on the consideration of feedback loops and evaluate the results in comparison to the previously analyzed methods (see Chapter 5).

5. Inclusion of Feedback Loops in Regulatory Activity Estimation

In all methods for inferring transcriptional activity previously described in Chapter 3 and 4, the concept of self-regulation via feedback loops (FBLs) has not been considered specifically, and we are not aware of any other method describing such a model. However, feedback loops are an important part of regulation processes in any cell type, enabling the regulation not only of gene expression but in turn also of TFs and other regulatory proteins [Brandman and Meyer, 2008].

In general, feedback occurs when the output of a system is passed back as input for the same system, forming a closed loop. Feedback can be either positive or negative, depending on how the respective values are referenced. Usually, positive feedback refers to the effect of self-reinforcement, tending to accelerate or intensify a process, whereas negative feedback describes a self-correcting behavior, slowing down a process and reducing the input signal. Further, a system can contain mixtures of positive and negative feedback where either positive or negative feedback can dominate [Ford, 2000].

In a cell, feedback loops are a common regulatory element in signaling and represent a central control mechanism driving cellular behavior [Sauro, 2017]. They enable the cell to mediate biological functions such as bistable switches or oscillations [Brandman and Meyer, 2008]. Feedback loops can contain either only one regulatory element (referred to as auto-regulation) or several elements, forming larger cycles. For example, the TF *PU.1* forms an auto-regulatory loop to control myeloid and early B-cell development [Leddin et al., 2011]. The mechanism of *ERK* regulation by *SHP2* forms a positive feedback loop that enhances the maintenance and invasiveness of breast tumors [Aceto et al., 2012]. In colorectal cancer cells, the prevalent signaling mechanism appears to be strong negative feedback from *ERK1/2* to *BRAF*, and the transcriptional activation of the *DUSP* family that inactivates *ERK* [Morkel et al., 2015] (see Figure 5.1).

In this work, we focus on transcriptional activity, as TF-gene interactions represent the predominant mechanism of gene regulation. TF loops occur in different recurring regulation patterns, called network motifs [Milo et al., 2002]. A TF with a positive (negative) auto-regulation describes a transcription factor that activates (represses) its own promoter. A TF binds to a gene, which in turn may encode a signaling molecule that plays a role in the cascades that regulate the TF's activity [Kel et al., 2019] (see Figure 5.2). Also, larger feedback loops are possible. In our graph model, a regulating TF is directly connected to its target gene, and vice versa, a protein producing gene to

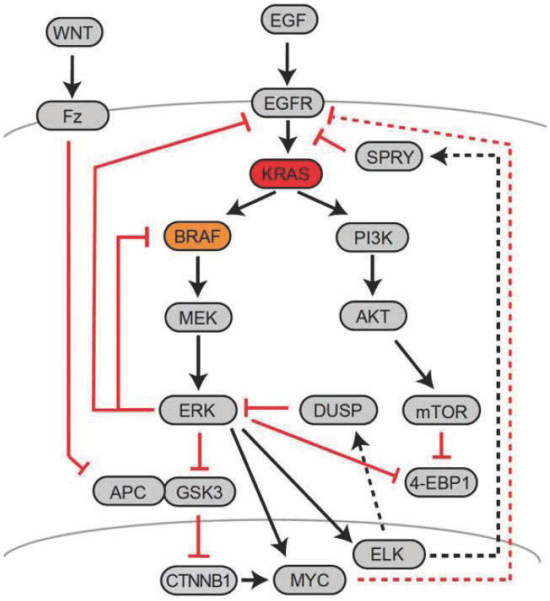


Figure 5.1.: Schematic representation major feedback mechanisms controlling MAPK (mitogen-activated protein kinase) activity in colorectal cancer (taken from [Morkel et al., 2015]). Major positive interactions are given as black arrows, while inhibitory interactions are given as red blocked lines. Solid lines indicate molecular interactions, whereas dotted lines indicate transcriptional control. Names frequently refer to a representative member of a multiprotein family.

the produced TF (see Figure 2.5). The gene-TF edges indicate that a gene produces a protein or signaling molecule, that plays a role in the cascades regulating the activity of the TF. The simplification of the biological process in the model omits the representation of the intermediate signaling molecule.

Our aim is to improve the currently in other methods for estimating TF activity presented models of cellular processes by taking into account the high importance of TF self-regulation. In this chapter, we propose a new method for estimating transcriptional activity which is the first one being able to consider feedback loops. We name our method Floræ (**F**eedback **lo**ops in **r**egulatory **a**ctivity **e**stimation)[1], as it is able to find loops in the regulatory network and subsequently attempts to infer the specific activity of TFs within the loop. Floræ implements an Expectation Maximization (EM) procedure, applying a Gaussian mixture model to the activity values of each TF in a feedback loop and the related gene expression values. Iteratively, we estimate the means and proportions of each Gaussian distribution of the mixture and use them to score TF

[1]We chose the name Floræ (old English spelling of the Latin plural of flora) due to its connection to biology and since the ligature æ graphically reminded us of an intertwined network.

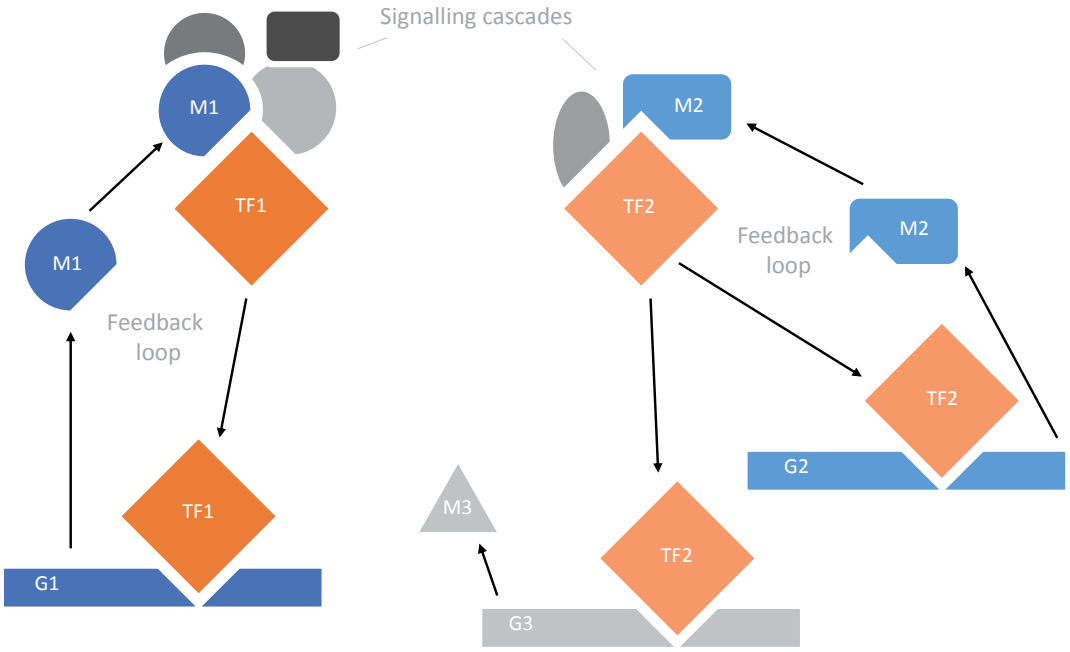


Figure 5.2.: Scheme of feedback loops in the gene regulatory network (adapted from [Kel et al., 2019]). The genes G1-G3 are controlled by TF1 respectively TF2. G1 and G2 encode for signaling molecules M1 and M2, that play a role in the cascades that regulate the activity of TF1 respectively TF2.

activity. To generate initial activity values and to assign activity values to TFs not part of a loop, we use the model from biRte [Fröhlich, 2015].

We evaluate the results of Floræ in comparison to the methods previously analyzed in this thesis, mostly based on synthetic data. We choose to use small artificial networks and to simulate according expression values since this approach represents a fast and integrated possibility to compare the results of different methods, while allowing us to control all parameters and to be able to interpret the results. We analyze the activity values inferred by biRte, RABIT, RACER and Floræ and examine the influence of the network's topology and the sample size on the results. Using Floræ, we are able to improve the identification of knockout and knockdown TFs in synthetic data sets. Additionally, we use the data sets presented in Chapter 4 to apply Floræ to real biological data. As expected, the results from Floræ were close to those from biRte, and only marginal improvements could be detected for the knockdown data from human and E. coli cell lines.

5.1. Method

5.1.1. Motivation

We propose Floræ (Feedback loops in regulatory activity estimation), a method for the inference of TF activity with a specific focus on the adequate analysis of feedback loops in the underlying gene regulatory network. A scheme summarizing input, model and output of the algorithm is given in Figure 5.3 (see also Chapter 3 regarding the notation). Floræ first finds loops in the regulatory network with a defined maximum cycle length. Regarding each specific loop, Floræ tries to find a consistent solution for the TF activities within the loop by alternating between two phases: the inference of TF activities and the inference of gene expression values. Hence, the estimations converge iteratively to a stable solution. In each phase, an Expectation Maximization (EM) algorithm [Dempster et al., 1977] with a Gaussian mixture model of two components is employed, reflecting active and inactive states of the TFs. The EM algorithm is an iterative process which maximizes the likelihood function of a parametric model in which some of the variables are "latent" (unknown) variables or treated as such [Balakrishnan et al., 2017]. Here, we use the EM algorithm as an optimization technique to numerically find the optimal parameters of the likelihood model, since it is analytically impossible or infeasible to directly calculate them. Using mRNA expression data of case and control samples and a TF-gene network as input, we use biRte to generate initial activity values and to assign activity values to TFs that are not comprised in a loop. BiRte uses Markov Chain Monte Carlo (MCMC) simulations to estimate TF activities. Both MCMC and EM are used to achieve the same goal here, which is solving the maximum likelihood estimation problem. The EM algorithm is typically used for the inference of point estimates as a simply understandable and easily implementable method, but has possibly a slow convergence [Rydén, 2008]. In comparison to biRte, Floræ performs an additional refinement of the solution of the global optimization problem by enforcing consistent values within the loops. Floræ eventually combines the estimations of biRte and its own TF activity values for the TFs comprised in a loop to score the activity of all TFs.

The choice of the network, which should be adapted to the available mRNA input data, can be handled by the user. We focus on TFs as they represent the predominant mechanism of gene regulation. Of course, our method could also be adapted to the analysis of other regulatory relationships, like miRNA-gene interactions. We will sketch the necessary adaptation in Section 5.3.3. In our current implementation of Floræ, we apply biRte to compute initial activity values for all TFs, which thereafter are used as starting values in the EM step. We choose biRte due to its good performance in our previous analyses (see Chapter 4), especially since it provides the best results when analyzing knockout and knockdown data. Since we analyze such experiments later on in simulation studies, we want to ascertain whether Floræ could even improve these results. Obviously, also other methods for estimating regulatory activity could be used for the computation of initial activity values, making Floræ adaptable to different applications and contexts.

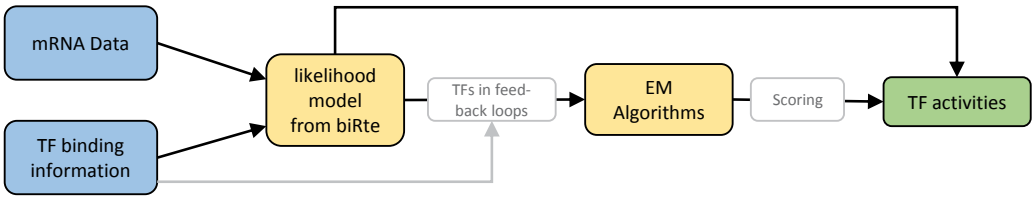


Figure 5.3.: Scheme of Floræ. The input data sets (marked in blue) are passed to biRte's likelihood model (yellow) which generates initial TF activity values. For all TFs included in a feedback loop, an EM algorithm (yellow) is used to score TF activities (green), others are taken directly from biRte.

5.1.2. Procedure

Floræ consist of four computational steps:

- Read the gene regulatory network and find all loops with the function `find_loops` (see pseudocode in Algorithm 1, Appendix A.6).
- Load the gene expression data together with the network and pass them to `apply_biRte` to get initial TF activity values (see pseudocode in Algorithm 2, Appendix A.6).
- Use the gene expression data, the loops and the initial activity values for all TFs comprised in a loop as input to the EM runs in function `EM_loops` (see pseudocode in Algorithm 3, Appendix A.6).
- Apply the `scoring` function (see pseudocode in Algorithm 4, Appendix A.6) to compute the final activities for the TFs comprised in a loop and assemble them with all other TF activity values from biRte for the TFs not included in a loop.

These steps are described in more detail in the following. We implemented Floræ in R, version 3.5.1 and use the packages igraph in version 1.2.2 and biRte (from Bioconductor) in version 1.16.0.

Loops in Graphs

As explained in Chapter 2, we use simple graphs to model the gene regulatory network provided as background knowledge to the inference methods. To apply Floræ, the detection of loops in a graph is the first necessary step. In graph theory, loops usually only refer to self-loops, i.e. edges that connect a node to itself. Here, we use a broader definition of the term "loop" and also include the notion of cycles. A cycle is a subset of the edge set of the graph that forms a directed path such that the first node of the path corresponds to the last.

To find all loops of length l, we search for subgraph-isomorphisms of the directed network with a ring of length l, i.e. a graph consisting only of a cycle of length l. An isomorphism is a bijection between two graphs G_1 and G_2 preserving their structure, i.e the image of any two adjacent vertices of G_1 are also adjacent in G_2 and vice versa [Gross and Yellen, 2003]. A subgraph-isomorphism is an isomorphism of a subgraph to another graph. We identify the subgraph-isomorphisms using the VF2 algorithm [Cordella et al., 2001] which is able to find such a mapping (if existing) between subgraphs of G_1 (the network) and G_2 (the ring). The depth-first algorithm starts with an empty mapping $M(k = 0)$. At each state k of the matching process, the algorithm computes the set of node pairs that are candidates to be added to the current state, which represents a partial mapping $M(k)$ between the two graphs. If a pair of vertices fulfills the subgraph-isomorphism condition, the mapping is extended and the algorithm is applied recursively. The algorithm explores all relevant mappings from G_2 to G_1 and returns all cycles of length l in the network. Since a cycle of length l has l different isomorphisms to itself, we remove such variations of the same cycle in the results of VF2.

In our implementation, we search only for loops of even length and until a maximal length of `max_length=4` (see Algorithm 1 in Appendix A.6). We do not search self loops ($l = 1$), since Floræ can only consider relationships of two different nodes in the network. The search for subgraph-isomorphisms is implemented in the `igraph` package in the function `graph.get.subisomorphisms.vf2`.

biRte

To calculate initial activity values for the subsequent EM algorithms, we use biRte [Fröhlich, 2015], see Algorithm 2 in Appendix A.6. Gene expression data for each sample, as well as the network is loaded as input. We use the function `birteRun` with the parameters `niter` and `nburnin` set to 10,000 and get an activity parameter for each TF and each sample as output.

EM Algorithms

For all TFs that are included in a loop, Floræ employs Expectation Maximization (EM) algorithms with Gaussian mixture models of two components to iteratively estimate means and proportions of each Gaussian distribution using the mRNA expression data. The theoretical background of the EM algorithm and the calculation of the parameters for the Gaussian mixture model has been described in Section 2.4.2.

From a high level view, two inference phases are alternated to represent the cyclic behavior of TF activity: first, we only consider those edges in the network pointing from a TF to a gene and use the gene expression data as observed variables to infer (the unobserved) TF activity values with an EM algorithm. Secondly, we reverse this attribution and use the inferred activity values as observed variables to infer a "gene activity" value, which we use as approximation for the gene expression, again using an

EM algorithm. Thus, in this second phase, we examine the directed edges in the network from a gene to TF. The estimated gene expression values are then again used as input to the first inference phase, and we iterate until the change of the parameters for the TF activity values is smaller than a certain threshold (see Algorithm 3 in Appendix A.6).

In more detail, we use the inferred TF activities from biRte to initialize θ and S in the first phase. Since S has to be a binary variable, we map the TF activities from biRte to 0 when $TF\_act < 0.5$ or to 1 in case $TF\_act \geq 0.5$ in each sample. If biRte assigns the samples to only one group, we randomly assign two samples to the other group. Using the measured gene expression values and the initialization, the EM algorithm iterates until convergence of Q and outputs an estimation for the group membership of the TF activities per sample (active or inactive, i.e. $S = 1$ or $S = 0$) in the first phase. These estimations are retrieved from the final values of $\tau_0(i) = P(S_i = 0|X_i = x, \theta)$, which reflect the probability that the TF was inactive in sample i. In the second phase, we simply reverse the attribution of measured and latent variables in the EM algorithm, meaning that we now consider the TF activity as measured and try to infer whether a gene was active or inactive, i.e. whether it produces a molecule for the signaling cascade of the TF or not. We use the inferred TF activities per sample from the first phase of the EM algorithm as values for X and assign the probability of $S = 0$ or $S = 1$ (the state of the gene) as gene activity. Note that we do not use the measured mRNA values for TF expression here, or compare the measured to the inferred values. Now, we obtain estimations for μ_s, σ_s and p, which represent the mean, variance and proportion of samples with active or inactive TFs. Iteratively, we now use the estimations for the gene expression, stored in τ_0, as input for another first EM phase where now the TF activities are the latent variables again. Thereby, we create an alternating procedure of two EM algorithms. The iteration terminates, when the change of the parameters μ_s, σ_s and p in a step where the genes are the latent variables is sufficiently small compared to the parameters of the previous run. When convergence is not achieved in less than 1000 iterations, we use different starting values for S in the affected EM run. In this case, one may restart `EM_loops` manually and use other initial values for S in the not converged EM run by switching the group assignment of one randomly chosen sample.

We compute TF activity values for all TFs included in a loop. However, a node can be part of multiple loops. In this case, we calculate the activity score for each loop separately and finally assign the highest score as overall TF activity, since we are interested in the existence of any high TF activity in a sample. In Floræ, we only explore loops with an even number of edges, since the algorithm is based on the notion of TF-gene and gene-TF interactions. At a TF-TF edge, both nodes would be initialized with values for S, making the mixture of Gaussian distributions inadequate. However, the algorithm could be extended to include also loops with TF-TF edges by introducing special calculation rules for those cases (see Section 5.3). We typically explore cycles of length two (TF_1 - $gene_1$ - TF_1) and four (TF_1 - $gene_1$ - TF_2 - $gene_2$ - TF_1) in our analyses.

Scoring

We use the final estimations of μ_s and p to score act_t, the activity of a TF t over all samples for the TFs in a loop. Since $p = P(S = 0)$ is a probability and μ_s is a mean of probability values, both parameters will lie in $[0, 1]$. If p is close to 0 or 1, meaning that the TF is active or inactive for most of the samples, we use the mean of the bigger group as estimate for the TF activity, which is either μ_0 or μ_1. If p is close to 0.5, signifying in about half of the samples the TF is active, and in the other half inactive, we suspect that the TF has a different behavior in the case and control group (although we do not check the accordance with the true group memberships). We therefore assign the absolute difference of μ_0 and μ_1 as TF activity value. Thus, we obtain high activity scores for TFs that are active in all samples and for TFs with high differential activity between two sample groups.

$$act_t = \begin{cases} \mu_1 & p \approx 0 \\ |\mu_1 - \mu_0| & p \approx 0.5 \\ \mu_0 & p \approx 1 \end{cases}$$

When these two cases (high activity in all samples and high differential activity) should be considered separately, the scoring heuristic could be adapted accordingly (see Section 5.3).

We use the sample size to determine the specific thresholds for p. We consider $p \approx 0$ respectively $p \approx 1$ when 75% of the samples (value rounded to a natural number) assign the TF an active respectively inactive state. In the other cases, when less that 75% of the samples are in one group, we consider $p \approx 0.5$. Thereby, we obtain the final scores for TFs included in a feedback loop. For the TFs, that are not included in a loop, we re-run biRte with the option single.sample=FALSE to get TF activity values over all samples, which lie in [0,1] as well. We finally assemble both results in a single list (see Algorithm 4, Appendix A.6).

5.2. Evaluation

We evaluate Floræ in comparison to biRte, RABIT and RACER (see Chapters 3 and 4) using different transcriptome data sets and networks. We mainly investigate synthetic data sets from five artificial networks, each including five feedback loops. With the tool GeneNetWeaver [Schaffter et al., 2011], we simulate knockout and knockdown experiments for each network and suppose that the highest change in activity will occur in the knocked out or knocked down TF when comparing case and control samples. By focusing on data with low complexity, we aim to thoroughly assess the results of Floræ and the influence of network topology and sample size. Additionally, we use the data sets presented in Chapter 4 to apply Floræ to real biological data.

Analogously to the evaluation in Chapter 4, we rank the absolute values of the computed TF activity scores, where the highest absolute activity value corresponds to rank 1. We appoint TFs that compare equal the same rank. Subsequently, a gap is left in the ranking numbers whose size is equal to the number of items that compare equal minus 1. Since activities equal to zero are not considered, the total number of ranked TFs can be different in each method and data set. We compute the ranks of all TFs, but evaluate only the rank of the TF that was knocked out or knocked down.

5.2.1. Synthetic Data

We simulate expression data using the tool GeneNetWeaver[2] [Schaffter et al., 2011]. Originally, the tool was developed for in silico benchmark generation and performance profiling of network inference methods. Here, we use GNW and five manually created gene regulatory networks to simulate expression data including the knockout respectively the knockdown of each TF in the networks.

Networks

The five artificial networks used for evaluation (see Figure 5.4), named network A to E, are relatively small to ensure a comprehensive interpretation of the results. They all comprise 10 TFs (denominated by Greek letters names), 24 genes (denominated by Latin letters) and 37 (networks A and D) or 38 (networks B, C and D) directed interactions (see also Table 5.1). We integrate two types of feedback loops:

- Direct loops (length two), where a TF influences the expression of a gene, which in turn expresses a protein affecting the expression or binding of the just mentioned TF, for example the path `Alpha - b - Alpha` in network A, and
- Indirect feedback loops (length four) including two TFs, where a first TF influences the expression of a gene expressing a protein affecting another TF, which in turn regulates a second gene and thereby again the first TF, for example the path `Beta - d - Gamma - e - Beta` in network A.

We include positive negative feedback loops, and a loop with mixed interactions in each network. Network A, B and C each contain two direct and three indirect loops, whereas the network D only contains direct and network E only indirect feedback loops. Of course, loops could include additional TFs and genes, but we aim to restrict the size and complexity of the model to be able to keep track of the network dynamics (see Section 5.3.3). The feedback loops are partly overlapping in the networks A, D and E, i.e. at least one TF is comprised in several feedback loops in these networks. In network B, all TFs within a loop are part of at least two loops, whereas in network C, the loops affect distinct TFs. Therefore, the number of TFs in a loop varies from network to network. The network's distribution of in-degree (the number of incoming edges) and out-degree (number of outgoing edges) are scale free, a property often observed in biological networks [Albert, 2005]. A summary of the network properties can be found in Table 5.1.

[2] http://gnw.sourceforge.net, accessed 10 September 2019

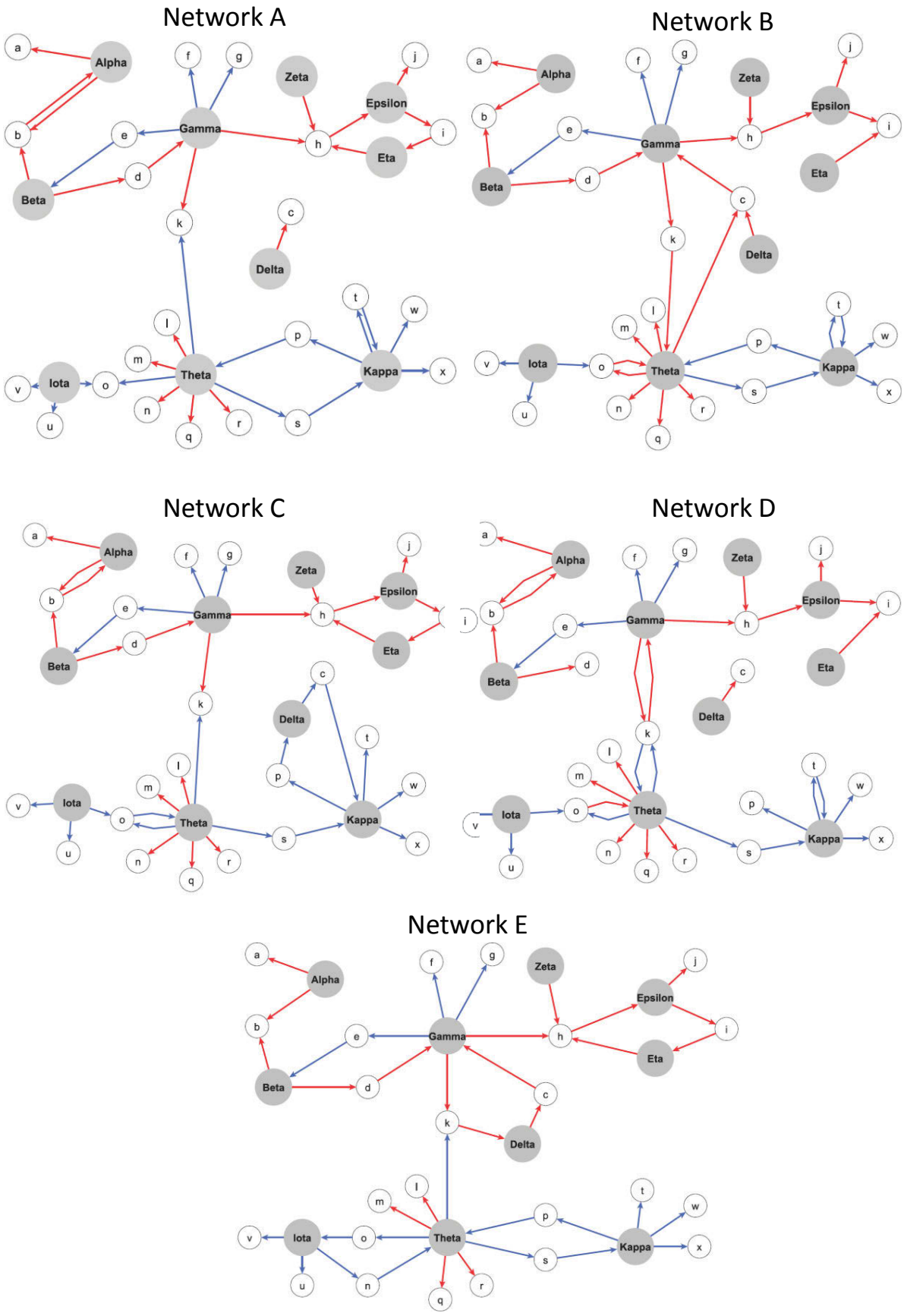


Figure 5.4.: Artificial gene regulatory networks A-E including feedback loops. Red (blue) arrows represent an inducing (repressing) effect of a TF on gene expression. TFs are labeled with Greek names, genes with Latin letters.

Property	Network A	B	C	D	E
number of TFs	10	10	10	10	10
number of genes	24	24	24	24	24
number of interactions	37	38	38	37	38
number of FBls	5	5	5	5	5
number of FBLs, length 2	2	2	2	5	0
number of FBLs, length 4	3	3	3	0	5
overlapping FBLs	partly	all	none	partly	partly
number of TFs within FBL	7	4	8	4	8

Table 5.1.: Characteristics of the artificial networks.

We further study the influence of randomized network edges by changing 10% or 50% of the interactions of network A. To this end, we randomly choose 4 respectively 19 network edges and assign them new connected nodes. In that process, we exclude edges that already existed in the original network. We generate 10 of these networks for each rate of randomization and evaluate the methods with each network, averaging the resulting ranks of knocked down and knocked out TFs at the end over all networks. Additionally, we evaluate network A without any feedback loops by eliminating the edges `b - Alpha`, `e - Beta`, `i - Eta`, `p - Theta` and `t - Kappa`.

Expression Data

In GNW, we simulate expression data using the artificial networks described above. GNW is able to endow a given network with dynamical models of gene regulation including both transcription and translation processes. It uses a thermodynamic approach accounting for both independent (additive) and synergistic (multiplicative) interactions [Schaffter et al., 2011]. The model further provides stochastic molecular noise and experimental noise observed in microarrays. The software can reproduce different types of in vivo experimental procedures:

- Wild type: Steady-state levels using the unperturbed network
- Knockout: Steady-state levels of single gene knockouts, providing an independent knockout for every gene of the network by setting the transcription rate of this gene to zero
- Knockdown: Steady-state levels of single gene knockdowns, simulating a knockdown of every gene of the network by reducing the transcription rate of the corresponding gene by half

GNW is a tool with a graphical user interface, available online as a web interface[3] or a downloadable stand-alone Java software[4], of which we use version 3.1 Beta for our

[3] http://gnw.sourceforge.net/genenetweaver.html, accessed 10 September 2019
[4] http://tschaffter.ch/projects/gnw/index.php, accessed 10 September 2019

analyses. For each network, we generate the expression data sets without removing auto-regulatory interactions and use deterministic ODEs as model. We omit the generation of dual knockouts and time series data. Using the same kinetic model, we generate a certain amount of samples per group: By default three samples, as this number reflects typical biological experiments, or five, ten and twenty samples to analyze the influence of sample size on the results. In each case, we analyze wild type (WT), knockout (KO) and knockdown (KD) data sets. We multiply all simulated mRNA concentrations with 100 to reach the range of biological array experiments and to circumvent numerical problems during TF activity estimation. Further, when the standard deviation within one experiment and sample group is zero, which is the case in some knockdown experiments for the KD TF, we add a small random error following a Gaussian distribution with mean zero and standard deviation 0.0001. For the results described subsequently, we run this expression data generation pipeline (see Figure 5.5, upper part in blue) 20 times per setting to obtain a distribution of the ranks of KO and KD TFs produced by activity estimation. PCA plots for an exemplary WT vs KO and WT vs KD data set based on network A are provided in the Appendix in Figure A.7, showing the separation of wild type and KO respectively KD samples. The WT samples are located closely together with the samples with `Alpha`, `Beta`, `Gamma`, `Delta` and `Zeta` KO/KD and separated from `Theta` and `Kappa` KO/ KD. The separation is more distinct in the KO compared to the KD plot.

5.2.2. Configuration

Using the artificial networks and the simulated expression data, we evaluate different methods for estimating TF activity, namely biRte, RACER and RABIT, and compare the results to the outcome of our own method, Floræ (see Figure 5.5, middle and lower parts). Based on our previous analyses of KO and KD data in Chapter 4, we do not include the method proposed by [Schacht et al., 2014] here, since its results were poor. Also, ISMARA [Balwierz et al., 2014] cannot be included in the analyses, as it can only be used with its own, proprietary underlying regulatory network. The evaluated methods were described in detail in Chapter 3. We apply the same configuration of RABIT, RACER and biRte as in Chapter 4, see page 52 for the specifications. For Floræ, we use our implementation, as described in Section 5.1.2.

5.2.3. Results

Effectiveness of Knockout and Knockdown

Next to WT, KO and KD expression data, GNW additionally outputs the original protein concentration which was used to simulate the expression data. This information can be used for the the evaluation of the effectiveness of KO and KD, as it reflects the true values of TF activity. We show the proportional change of protein concentration for all simulated expression data sets in Figures 5.6 (KO) and 5.7 (KD). When comparing the TF activity of WT and KO samples, we notice, as expected, that the knocked out TF has always the lowest protein concentration of all TFs, and the variability of the KO

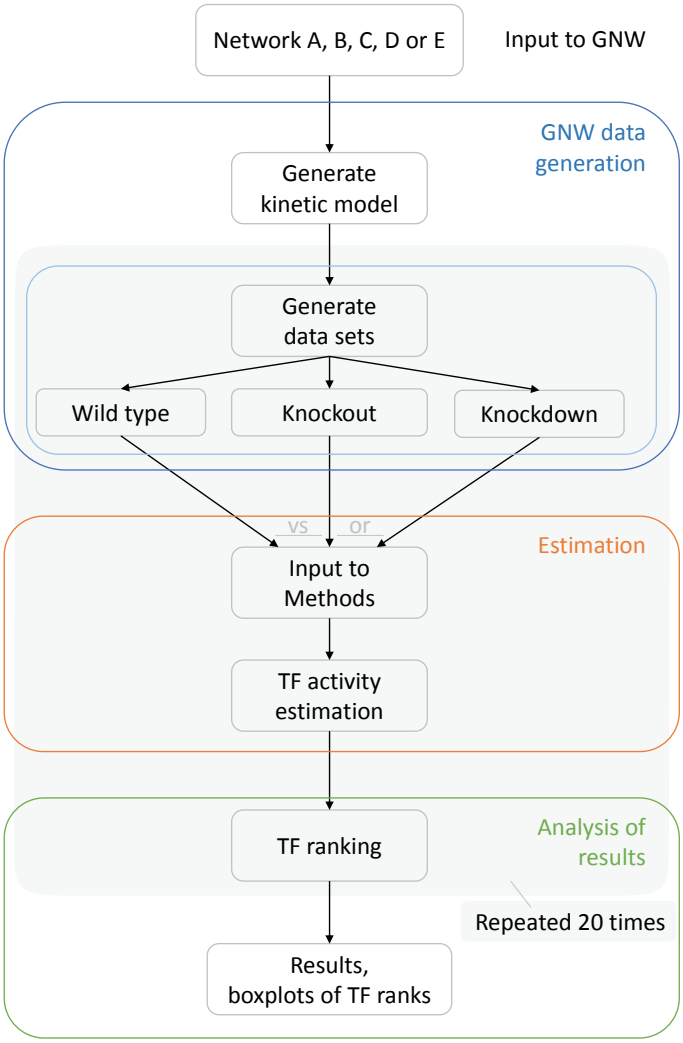


Figure 5.5.: Standard pipeline for data generation using GNW to simulate expression data, TF activity estimation and analysis of the results.

TF's activity over all samples is very small. For the KD experiments, the results are less clear. Although the protein concentration of the KD TFs is, as expected, approximately divided in half, other TFs show sometimes equal proportional changes. For example in the KD of `Eta`, also `Epsilon` has equally low activity values on average. Therefore, it should be a much easier task for the methods to determine the KO TF compared to the identification of the KD TF. Further, the protein concentrations of the TFs within the positive feedback loops, like `Alpha` and `Eta`, have a high variability in nearly all KO and KD experiments. Still, the median activity of all non-KO and non-KD TFs are close to zero, meaning that on average their concentration is not changed when comparing KO/ KD and wild type samples.

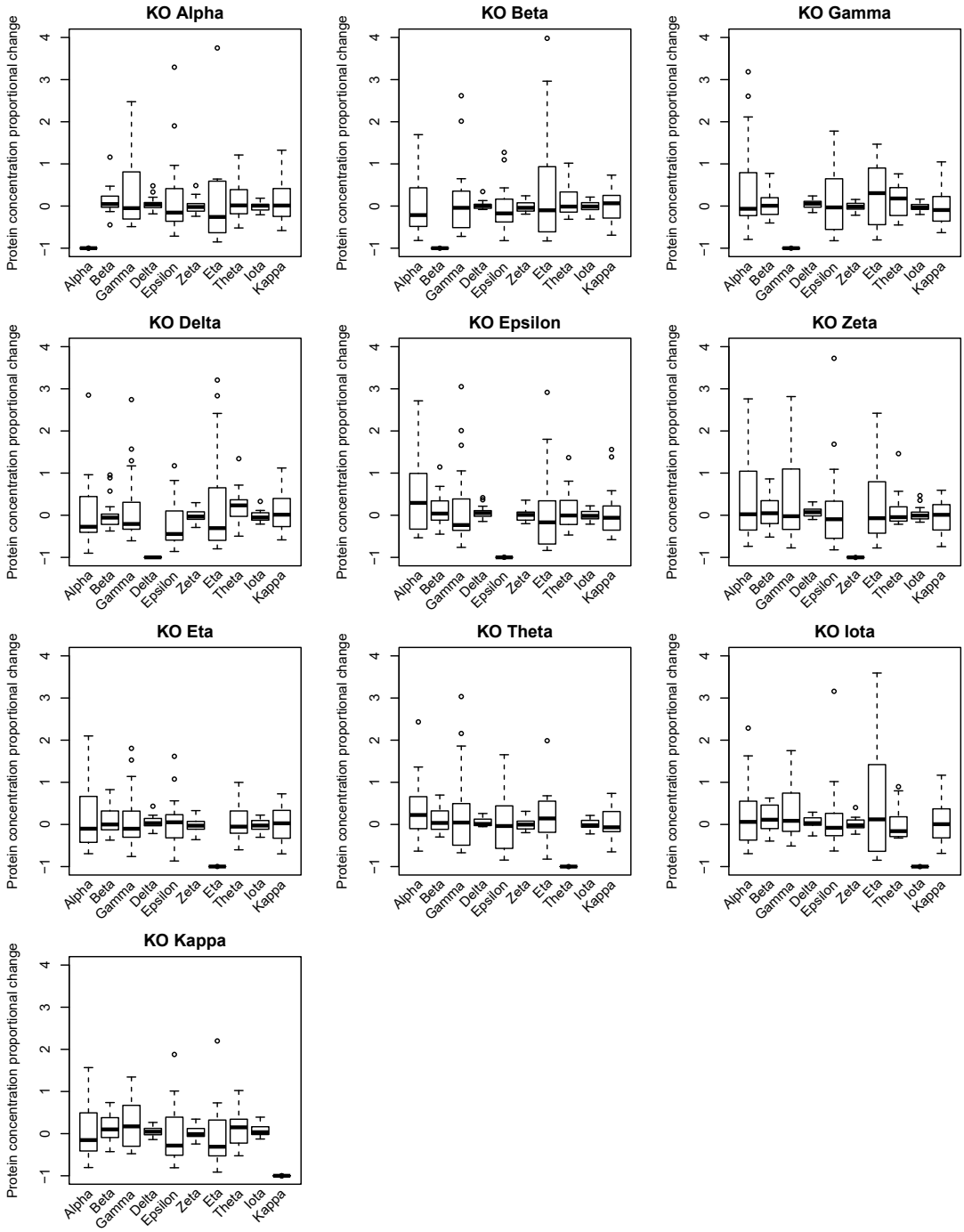


Figure 5.6.: Boxplots of the relative changes of protein concentration of all TFs given by GNW of WT vs KO samples. The change's median is represented by a bold line, the boxes range from 25^{th} to 75^{th} percentile, representing the interquartile range. Each plot shows a KO experiment, the heading indicates the corresponding KO TF.

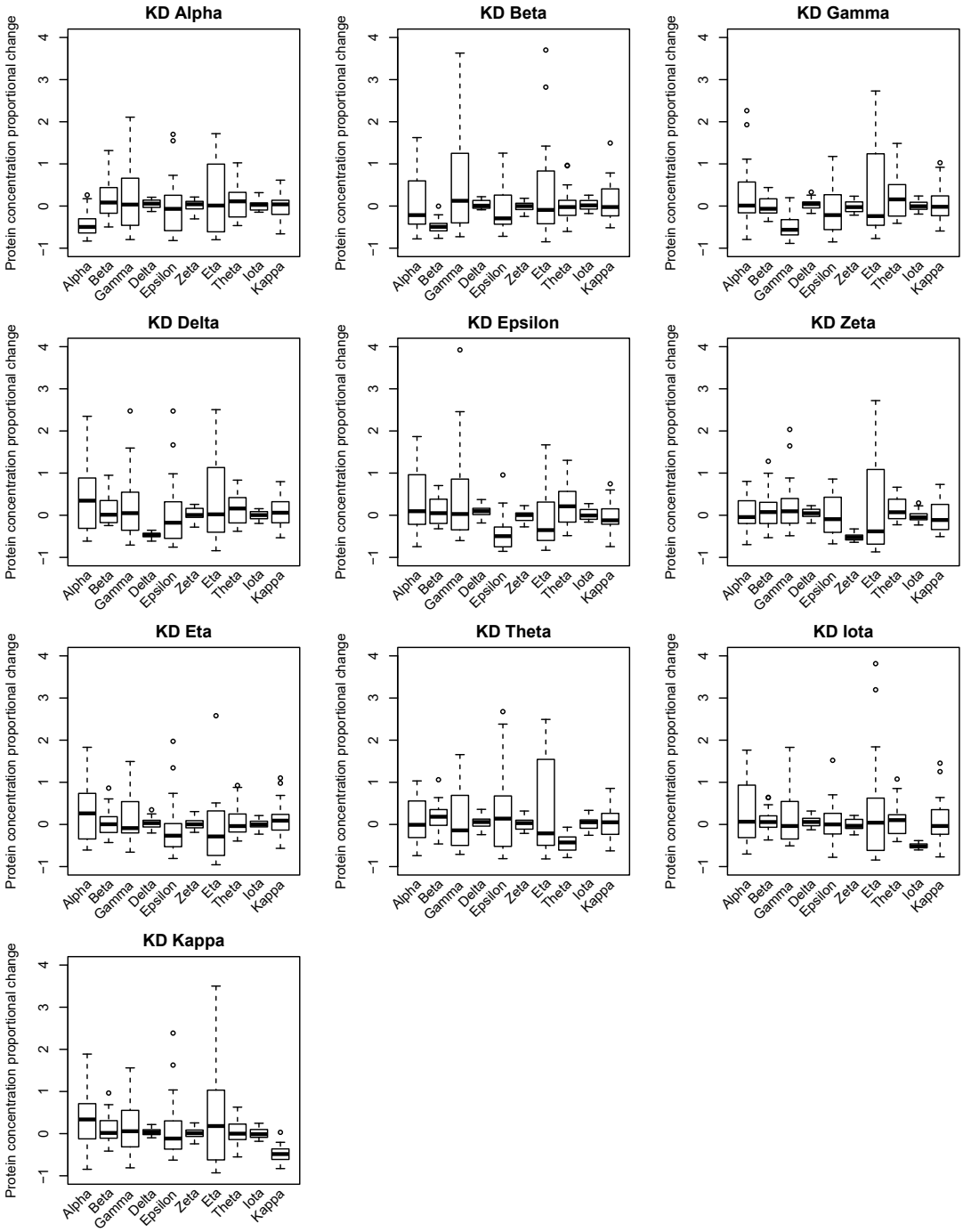


Figure 5.7.: Boxplots of the relative changes of protein concentration of all TFs given by GNW of WT vs KD samples. The change's median is represented by a bold line, the boxes range from 25^{th} to 75^{th} percentile, representing the interquartile range. Each plot shows a KD experiment, the heading indicates the corresponding KD TF.

Overview of Ranking of KO and KD TFs

We use all five artificial networks described in Figure 5.4 and the according simulated expression data for wild-type, knockout and knockdown samples (three samples per experiment) to compare the resulting TF activity ranks of Floræ, biRte, RABIT and RACER. Overall, when comparing the number of KO TFs ranked on position 1 or 2 of all methods (median rank), Floræ is able to improve the identification of KO TFs in four of the five networks (A, B, C, E) and yields equally good results as biRte and RABIT for the network D (see Table 5.2). In the KD data sets, Floræ yields in all five networks the best results, together with RABIT (networks B and C). However, for both KO and KD data sets, the median ranks of RABIT are partially based on less data points compared to the other methods, since RABIT does not provide a rank for the KO or KD TF in all 20 data sets. Therefore, the results of RABIT are as good as those from Floræ in some cases, but are less reliable. The number of correctly identified KO TFs by Floræ range from 6 (network D) to 10 (network B) and from 6 (networks A, C) to 8 (networks B, D, E) for the KD TFs. On average over all networks, Floræ is able to identify 8 out of 10 KO TFs and 7 out of 10 KD TFs.

		Network				
Data type	Method	A	B	C	D	E
Knockout	Floræ	8	10	9	6	7
	biRte	5	7	8	6	6
	RABIT*	5	9	7	6	5
	RACER	0	3	2	0	2
Knockdown	Floræ	6	8	6	8	8
	biRte	5	6	1	7	5
	RABIT*	5	8	6	7	6
	RACER	3	2	3	1	3

Table 5.2.: Number of KO and KD TFs ranked on position 1 or 2 by each method (median ranks) for each network. The best method per network and data type is marked in green. RABIT (marked with an asterisk) partly does not provide any ranking of the KO or KD TF, the median is calculated on the available ranks and does not consider the number of missing values.

The detailed results per network are shown in Figure 5.10 (network A, WT vs KO), Figure 5.11 (network A, WT vs KD) and in the Appendix A.9 (networks B-E, both KO and KD). For each network, these figures show boxplots of the resulting ranks per method and per knocked out respectively knocked down TF, aggregating the results of 20 runs of data generation and TF ranking. We will describe the results for network A in detail in the next section. Here, we present the aggregated results over all TFs and networks, which are shown in Figure 5.8 (WT vs KO) and 5.9 (WT vs KD). The results per KO and KD TF over all networks are shown in Appendix A.8.

When comparing the results by differentiating whether a TF was part of a loop or not, Floræ yields particularly good results for the identification of KO and KD TFs that are comprised in one ore several feedback loops (see Figures 5.8 and 5.9). Floræ ranked 24 out of 33 KO TFs that are part of a loop on position 1 or 2, whereas biRte and RABIT are only in 19 and RACER in 8 cases able to identify the KO TF. For KO TFs not part of a loop, the results from Floræ, biRte and RABIT are nearly equally good (Floræ: 16, biRte 15, RACER 13 KO TFs on rank 1 or 2 out of 17). RACER only identified 3 KO TFs here. We find comparable results also in the KD data sets, where Floræ ranked 25 out of 33 KD TF within a loop first or second (biRte: 14, RABIT: 22, RACER: 8) and 11 out of 17 KD TFs (biRte and RABIT: 10, RACER: 5). Overall, we observe that Floræ ameliorates the identification especially of those KO and KD TFs that are included in a feedback loop, which reflects that Floræ achieves the goal of improving the estimation of regulatory activity in synthetic data sets.

Ranking of KO and KD TFs in network A

We now analyze the results of Floræ, biRte, RABIT and RACER for network A described in Figure 5.4 and the according simulated expression data for wild-type, knockout and knockdown samples (three samples per experiment) in more detail as an example of how Floræ can improve the identification of KO and KD TFs. The results are shown in Figure 5.10 for the WT vs KO data, and in Figure 5.11 for the comparison of WT vs KD. The figures show boxplots of the resulting ranks per method and per knocked out respectively knocked down TF, aggregating the results of 20 runs of data generation and TF ranking.

For the KO data, Floræ, biRte and RABIT rank the KO TF in four out of ten cases first (median rank), attributing the highest change of TF activity to the actual KO TF and thus identifying the KO TF correctly. When looking at the top 2 TFs, Floræ ranks on average eight out of ten TFs on the first or second position, followed by RABIT and biRte (both five KO TFs). Floræ has in nearly all KO scenarios comparable or better results than biRte or RABIT, but due to stringent filtering thresholds in RABIT, often no activity score was assigned to the KO TF. Hence, the boxplots of RABIT are in some cases (marked with an asterisk in the label of the according plot) based on three to six values only, compared to 20 values for the other methods. RACER does not rank any KO TF into the top 2 and is therefore clearly outperformed by the three other methods. The methods (except for RACER) yield particular good results for the KO of `Delta`, `Epsilon`, `Iota` and `Kappa`. From these TFs, fewer genes and fewer TFs are reachable in network A, compared to more central TFs like `Beta`, `Gamma` and `Theta`, hinting at a relatively easy optimization procedure in these cases. The results for the nodes comprised in a feedback loop (`Beta`, `Gamma`, `Epsilon`, `Eta`, `Theta` and `Kappa`) are particularly good for the method Floræ, as for all these TFs (except `Gamma`) the median rank is one or two. Especially for the central TFs in network A, `Beta` and `Theta`, Floræ provides much better ranks compared to the other methods. The variation of TF KO ranks over different data generation and TF activity estimation runs is quite small, and not dependent on the specific KO TF or method. Typically, the interquartile range, representing 50% of the results, covers only two ranks.

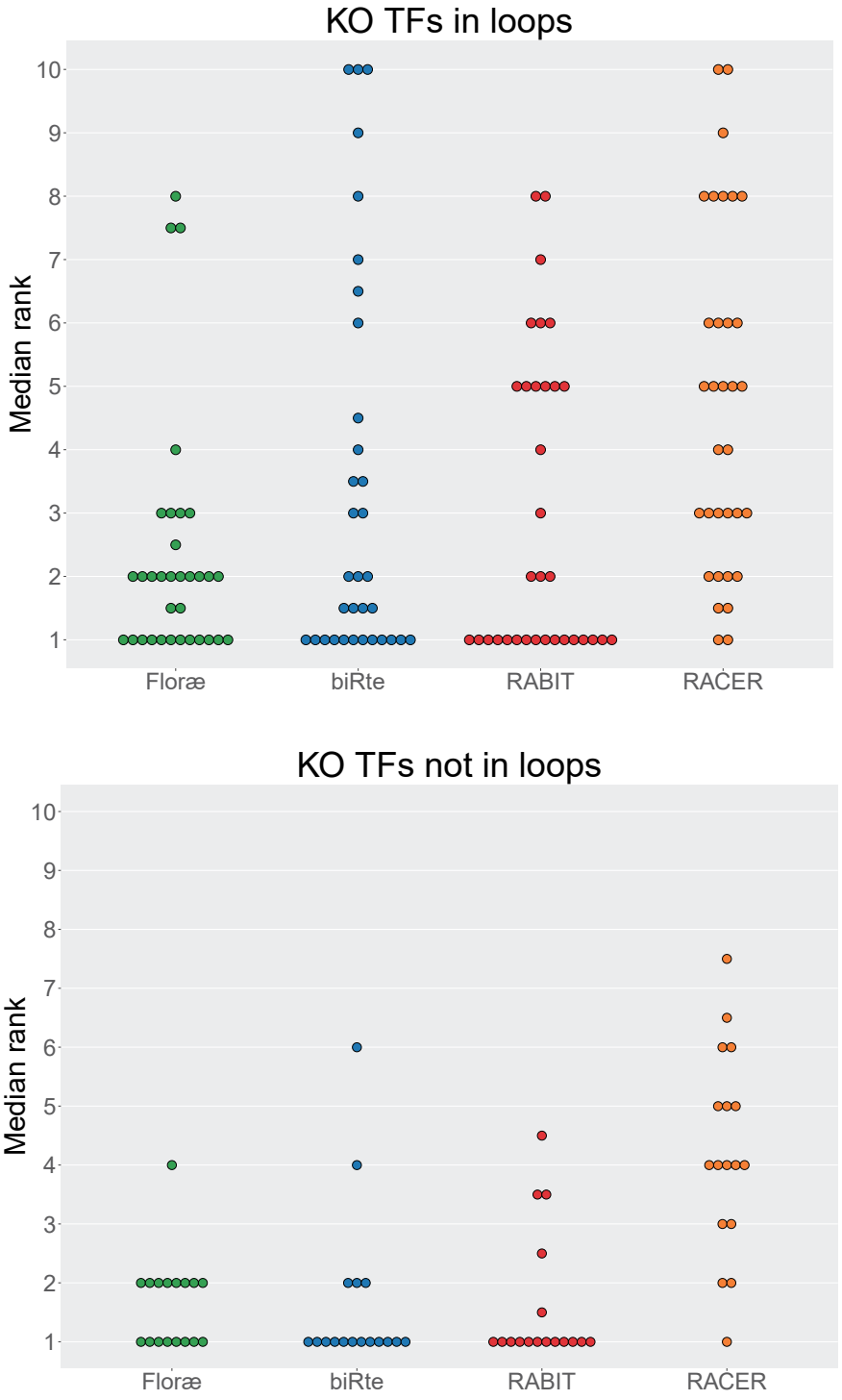


Figure 5.8.: Median ranks over all networks and KO TFs of Floræ (green), biRte (blue), RABIT (red) and RACER (orange). The upper plot contains the median ranks for all KO TFs, that are comprised in one or several loops in any network, whereas the lower plot shoes the median ranks for all other KO TFs.

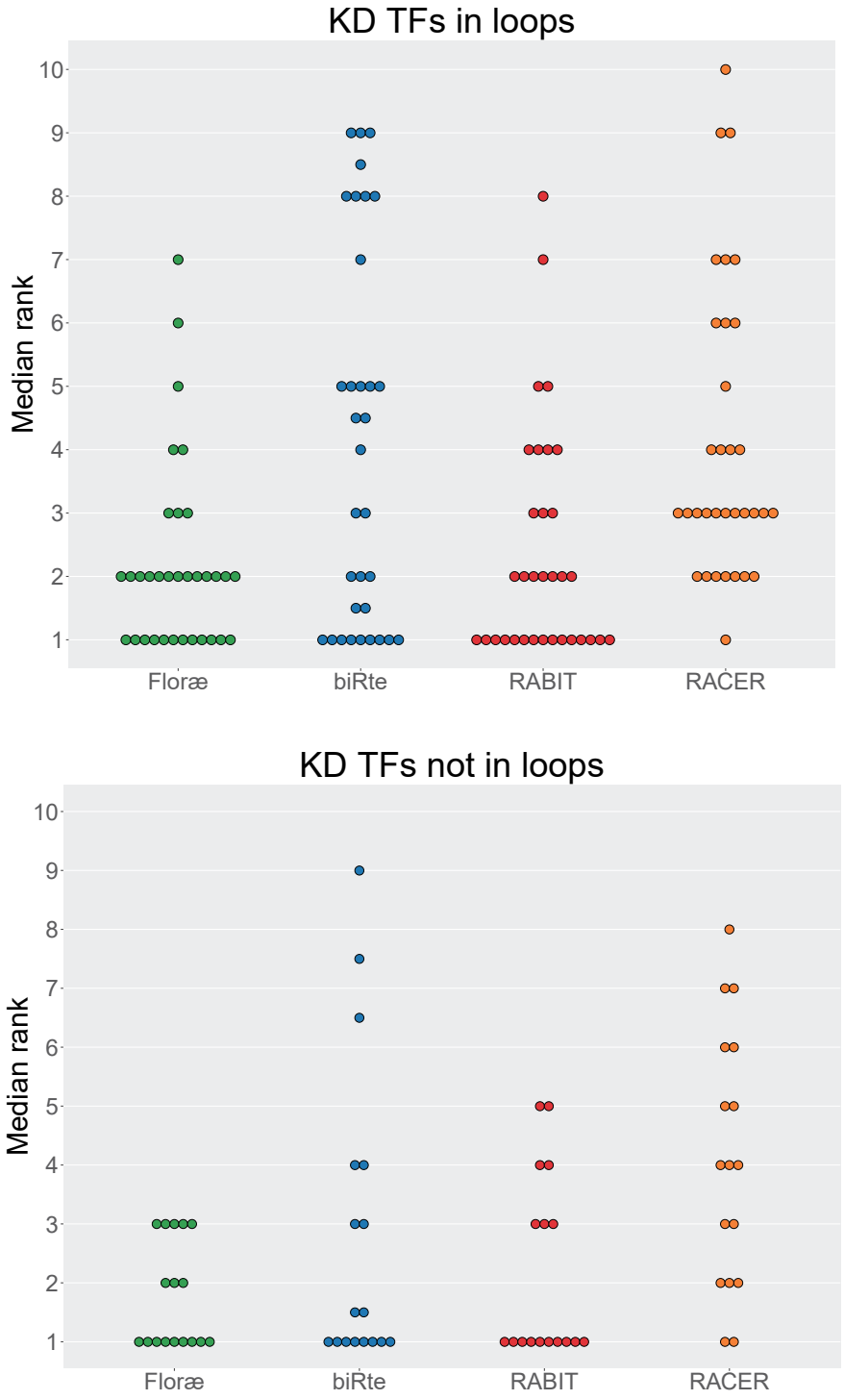


Figure 5.9.: Median ranks over all networks and KD TFs of Floræ (green), biRte (blue), RABIT (red) and RACER (orange). The upper plot contains the median ranks for all KD TFs, that are comprised in one or several loops in any network, whereas the lower plot shoes the median ranks for all other KD TFs.

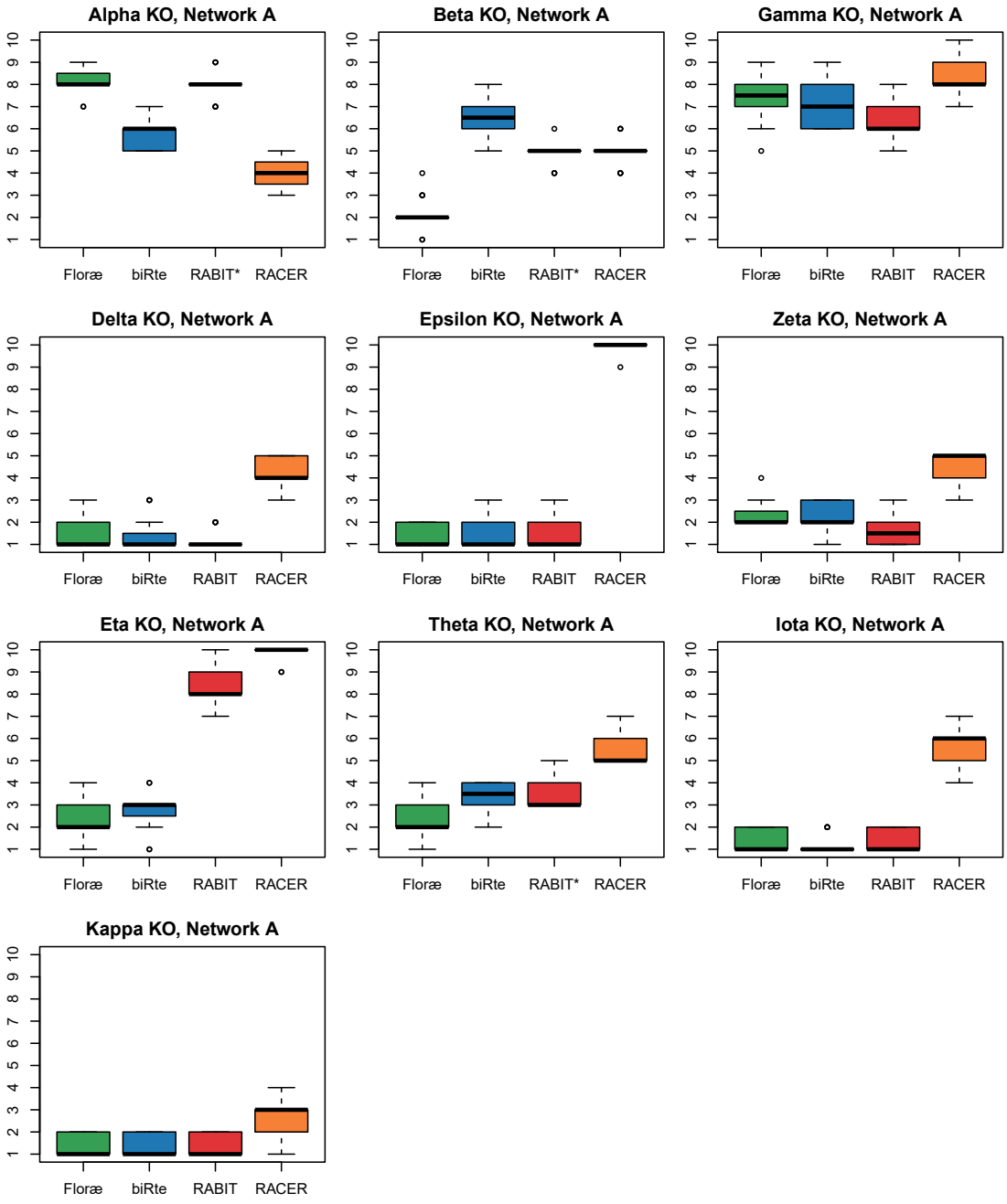


Figure 5.10.: Boxplots showing the TF activity ranks of Floræ (green), biRte (blue), RABIT (red) and RACER (orange) for all ten TF KOs based on network A, 20 runs of data generation and TF ranking. Median ranks are represented by a bold line, the colored box ranges from 25^{th} to 75^{th} percentile, representing the interquartile range. See main text for RABIT*.

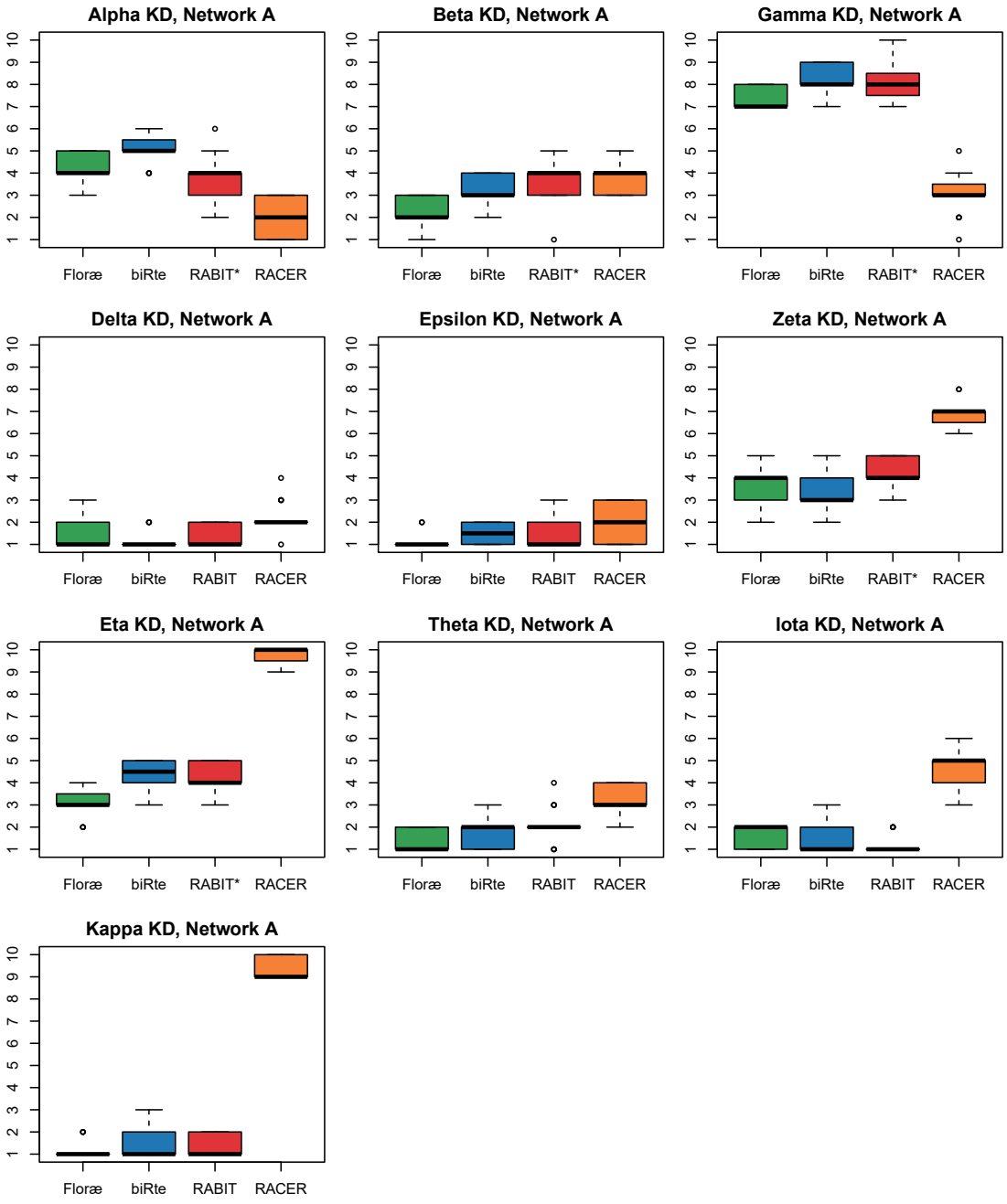


Figure 5.11.: Boxplots showing the TF activity ranks of Floræ (green), biRte (blue), RABIT (red) and RACER (orange) for all ten TF KDs based on network A, 20 runs of data generation and TF ranking. Median ranks are represented by a bold line, the colored box ranges from 25^{th} to 75^{th} percentile, representing the interquartile range. See main text for RABIT*.

The results for the KD data are comparable. Floræ, biRte and RABIT are able to detect the KD TF, i.e. assigning the first rank to the KD TF in four cases (Floræ and RABIT), respectively three cases (biRte). When looking at the top 2 TFs, Floræ ranks on average 6 out of 10 TFs on the first or second position, followed by RABIT and biRte (both 5 KD TFs) and RACER (3). Like in the KO scenario, RABIT assigns only three to six times a rank to the KD TF in half of the KD experiments, thus limiting the explanatory power of the according boxplots (marked with asterisk). Again, the TFs `Delta`, `Epsilon`, `Iota` and `Kappa` in network A yield quite good results in Floræ, biRte and RABIT, being the top 1 or top 2 TF on average. Also, Floræ performs well on the TFs comprised in a feedback loop and has better or comparable results to biRte. Surprisingly, the performance of all methods does not considerably decrease in the KD scenario compared to the KO data sets, even when the highest proportional change of protein concentration from WT to KD did not affect the KD TF itself but another TF, as described before.

Effect of Feedback Loops

To assess the influence of feedback loops in the network on the performance of the methods, we evaluate the methods using a network without any feedback loops by eliminating the edges `b - Alpha`, `e - Beta`, `i - Eta`, `p - Theta` and `t - Kappa` in network A. As expected, Floræ now yields similar results like biRte (see Table 5.3), due to their methodological resemblance. Note that the results are not identical, since we average the results from biRte over 100 runs (see Section 4.1), but Floræ uses only one initialization of biRte. Overall, the exclusion of feedback loops only diminishes the performance of Floræ in terms of the detection of KO or KD TFs, whereas the results of RABIT remain nearly unchanged. The ranks inferred by RACER change a lot when changing the underlying regulatory network, both positively and negatively. Without the inclusion of feedback loops, Floræ loses its capacity to yield much better results compared to biRte for the KO and KD of `Beta`, and the marginally better ranks for `Eta` (KO and KD), `Zeta` (KO) and `Alpha` (KD), as expected.

Sample number

We further analyze the influence of the sample size on resulting TF activity ranks in network A. In our first analysis (see previous two sections), we used three samples per wild-type and knockout respectively knockdown experiment, as biological experiments tend to provide only a small number of samples per group. However, we are interested whether a higher number of samples could improve the inference of TF activity. The results for KO are presented in Figure 5.12 and for KD in the Appendix in A.10. The plots indicate mean ranks and the according standard error of the mean (SEM) calculated by $SEM = \frac{\sigma}{\sqrt{n}}$ with σ the sample standard deviation and n the sample size. In general, the sample size has no crucial impact on the estimated TF activity ranks and the results are mainly dependent on the actual KO TF and the inference method. KO TFs that already had good results for a sample size of three are the top 1 or 2 TF also

TF	Floræ		biRte		RABIT		RACER	
	FBL	no FBL	FBL	no FBL	FBL	no FBL	FBL	no FBL
Alpha KO	8	8	6	9	8	6	4	3
Beta KO	2	6	6.5	6	5	5	5	9
Gamma KO	7.5	8	7	9	6	6	8	9
Delta KO	1	1	1	1	1	1	4	1
Epsilon KO	1	1	1	1	1	2	10	2
Zeta KO	2	3.5	2	3	1.5	2	5	5
Eta KO	2	4	3	4	8	8	10	10
Theta KO	2	1.5	3.5	1.5	3	1	5	8
Iota KO	1	1	1	1	1	1	6	5
Kappa KO	1	1	1	2	1	2	3	8
Alpha KD	4	6	5	7	4	4	2	3
Beta KD	2	5	3	5	4	4	4	5
Gamma KD	7	8	8	9	8	8	3	6
Delta KD	1	2	1	3	1	2	2	2
Epsilon KD	1	1	1.5	1	1	2	2	2
Zeta KD	4	3.5	3	3	4	4	7	10
Eta KD	3	4	4.5	4	4	4	10	7
Theta KD	1	1	2	1	2	2	3	4
Iota KD	2	1.5	1	1	1	1	5	7
Kappa KD	1	1	1	1.5	1	1	9	7

Table 5.3.: Effect of the inclusion of feedback loops (FBL) in the underlying regulatory network. The table shows the median ranks of KO and KD TFs for all methods based on 20 runs of data generation and TF activity estimation. Improved ranks with the use of the network with FBLs are colored in green (improvement > 1 rank), worse ranks in red.

for higher sample numbers, for example `Delta`, `Epsilon`, `Iota` and `Kappa`. For other TFs, like `Zeta` and `Eta`, the increase of sample size does not have an improving effect on identifying the KO TF. Only `Beta`, `Theta` and partly `Gamma`, TFs with feedback loops located centrally in network A , show better ranks for higher sample sizes (except ranks inferred by RACER in `Theta` and biRte and RABIT in `Gamma`). In general, the results of RACER are contrary to the results of the other methods and show the highest variability across different sample sizes. For example, RACER provides small ranks for the `Alpha` KO using 5 and 20 samples, which no other method is able to yield, but is not able to detect the `Theta` or `Kappa` KO for sample sizes larger than 5, whereas the other methods obtain quite good results. Floræ is for three KO TFs (`Beta`, `Gamma` and `Eta`) the best method, i.e. achieves lowest ranks for all sample sizes. In 5 other cases (`Delta`, `Epsilon`, `Theta`, `Iota` and `Kappa`), Floræ yields comparable results to biRte and/ or RABIT. Regarding the KD data sets, comparable results are achieved (see A.10).

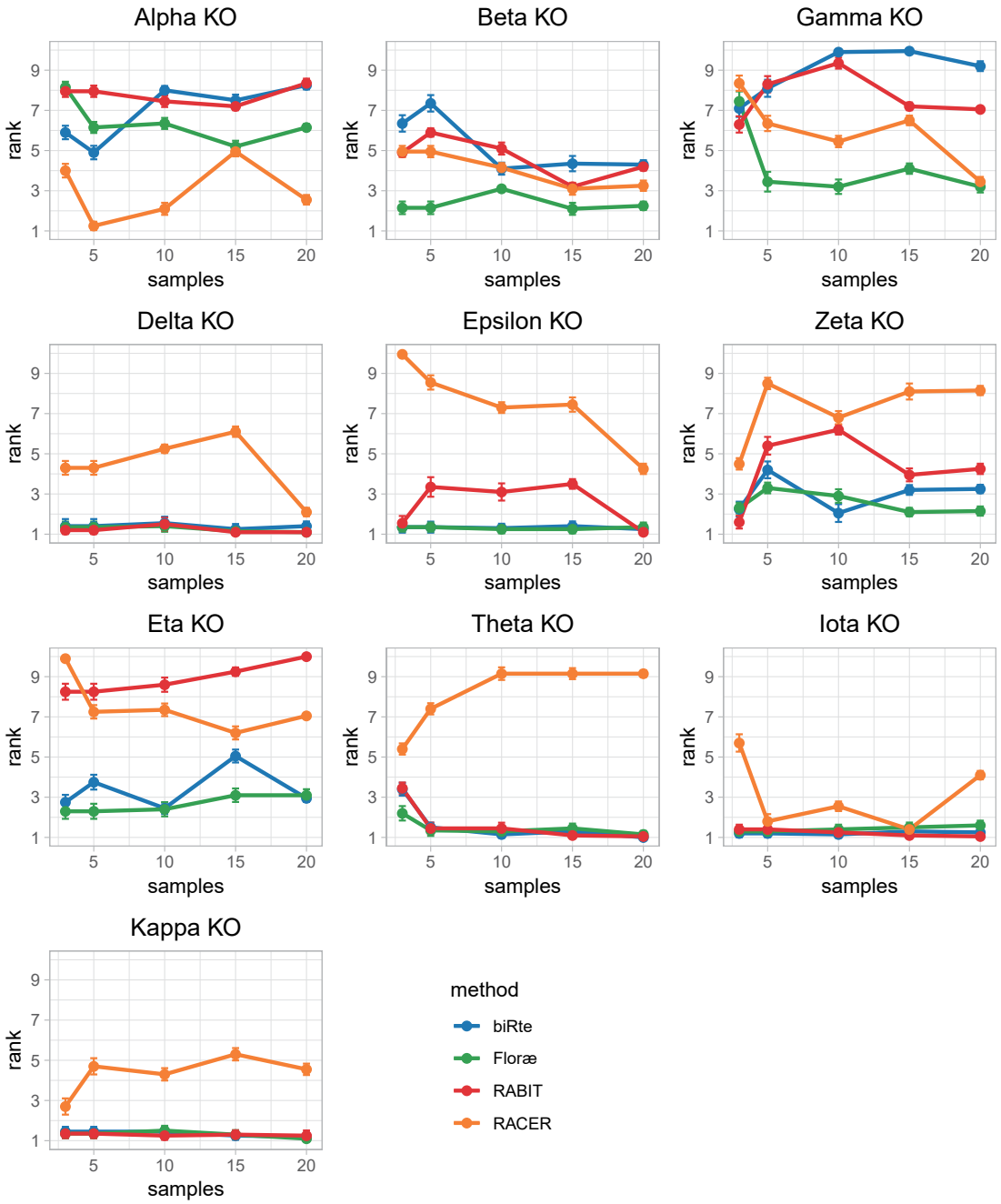


Figure 5.12.: Mean ranks and according standard errors of the mean of TF activity ranks for all ten knockout TFs using a varying number of samples (3, 5, 10, 15 and 20) for both wild-type and knockout experiments. Per sample size, TF activity ranks are calculated on the basis of network A, 20 runs of data generation and TF ranking using biRte (blue line), Floræ (green), RABIT (red) and RACER (orange).

Network Randomization

In our previous analyses, we used the artificial regulatory networks employed for data generation also as input to the methods for estimating TF activity. However, in biological applications, the underlying gene regulatory network and its structure are almost never known completely, even in simple organisms. We therefore investigate the influence of network changes on TF activity estimation. After data generation, we randomize the edges from network A by changing either 10% or 50% of the interactions, randomly removing 4 respectively 19 network edges and inserting new TF-gene edges with random direction and effect (enhancement or repression) elsewhere. We generate 10 of these networks for each rate of randomization and exclude edges that already existed in the original network. We do not apply a degree-preserving randomization (such as rewiring existing edges), since in a biological application the node degrees of the (partly) known underlying regulatory network do not have to be correct. Further, during randomization, we do not preserve the originally present feedback loops of the artificial network A or prevent the formation of new ones. This led to a decrease of the number of feedback loops for the networks with 50% changed interactions. The remaining number of loops per randomization rate is indicated in Appendix A.11, as well as an exemplary randomized network for each randomization rate.

We evaluate the methods using each randomized network as input, as well as the same synthetic WT, KO and KD data sets already used before, with three samples in each group. We average the resulting ranks of knocked down and knocked out TFs over all randomized networks. The results show, as expected, a higher variability across the ten different randomized networks, compared to the use of the original artificial network A. Especially the results from RACER are highly dependent on the actual network provided as input, generating unstable TF activity ranks. When comparing WT and KO samples with a network randomization rate of 10% (see Figure 5.13), Floræ, biRte and RABIT are still capable of identifying those KO TFs that already had good results with the original artificial network A (`Delta`, `Epsilon` and `Iota`). Floræ ranks seven of the ten KO TFs on rank 1 or 2 (median rank), followed by RABIT (six) and biRte (four). The relatively small variance of the ranks provided by RABIT is partly caused by a high number of missing values, not assigning any rank to the KO TF and thus limiting the explanatory power of the according boxplots (marked with asterisk). When changing 50% of the network edges (see Appendix A.11), the variability of the resulting TF ranks is much higher, also for Floræ and biRte. RABIT provides good results, but overall rarely ranks the KO TF at all, leading to unstable results which are highly dependent on the actual network and samples provided as input. Floræ, biRte and RABIT yield good results for the KO of `Iota`, and Floræ still identifies `Delta`, `Epsilon` and `Kappa` as KO TF on rank 1 to 3 (median rank). The results from the comparison WT vs KD are comparable: for 10% randomization rate, Floræ and RABIT rank five out of ten TFs on rank one or two (`Delta`, `Epsilon`, `Theta`, `Iota` and `Kappa`), followed by biRte (three) and RACER (zero). When 50% of the network edges are rewired, the results for `Epsilon`, `Theta`, `Iota` and `Kappa` are still acceptable, being ranked on average at first

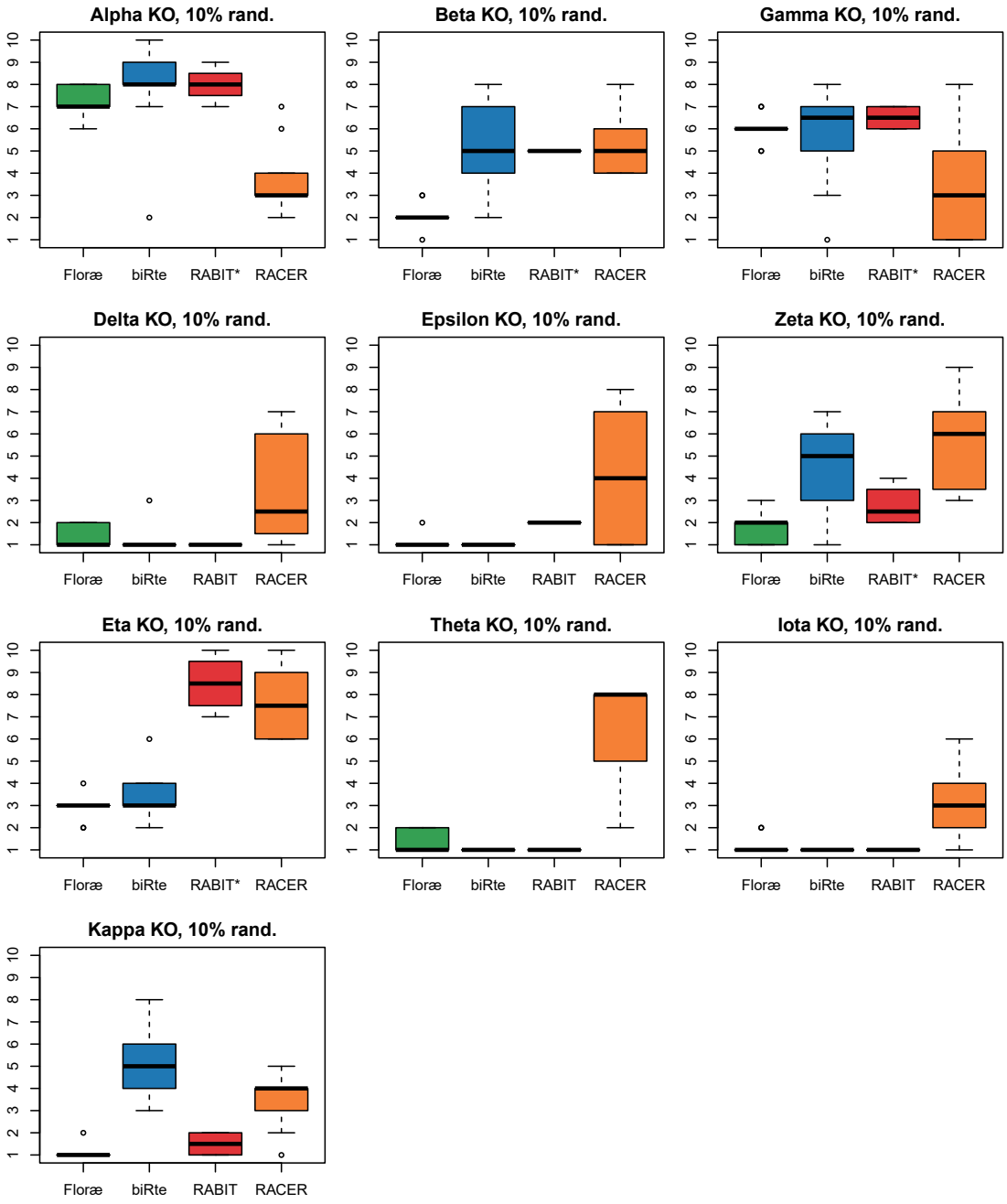


Figure 5.13.: Effect of network randomization of network A (10%), WT vs KO samples. Boxplots showing the TF activity ranks of Floræ (green), biRte (blue), RABIT (red) and RACER (orange) for all ten TF knockouts.

or second position by Floræ, biRte and RABIT. However, the variability of the results from all methods is much higher compared to the original network A or the networks with 10% randomization rate. In nearly all KD TFs, RABIT does not provide ranks for all randomized networks, leading to a high number of missing results, which skews the results. The results for the WT vs KD comparison can be found in Appendix A.11.

Application to biological data sets

We briefly evaluate the application of Floræ to real biological data sets. We use a part of the TCGA data, as well as the knockdown and knockout experiments described in Chapter 4. We first analyze the presence of loops in the human text mining and E. coli regulatory network. Whereas the number of loops in the network from E. coli is relatively small (107 self-loops, 12 loops each of length two and three, one loop with four nodes and no loops of length five to ten), the text mining network has a lot more loops (50 self-loops, 30 loops of length two, 30 of length three, 67 of length four, 170 of length five and 16.298 of length ten). This high number of long cycles occurs for example when in a cycle of length ten, only one node changes compared to another cycle of the same length and the other nine nodes are identical. Since we set the maximum loop length in Floræ to four, we expect that the results for human data will change more than the results of E. coli.

We select the COAD (colon adenocarcinoma) mRNA data from TCGA, described in Chapter 4, to compare the top 10 ranked TFs from biRte, RABIT, RACER and Floræ. Like before, we use the text mining network as TF-binding information. The results are given in Table 5.4. Overall, the results from Floræ have a large overlap with the results from biRte (seven TFs). In the top 10 TFs of all methods, seven TFs are found by two methods and another three TFs by three methods. This overlap is larger compared to the results from Chapter 4, since we do not include the method by [Schacht et al., 2014] here, which had no overlap in the top10 of the COAD data set (see Table 4.1). Further, we find *PHOX2B* in the top 10 of all four methods. *PHOX2B* was found to be hypermethylated in colorectal cancer and might be used as biomarker in early diagnosis [Li et al., 2012]. Further, *PHOX2B* is susceptible to play a role in Crohn's Disease [Lauriola et al., 2011] and other diseases, like neuroblastoma [Di Zanni et al., 2017; Yin et al., 2016]. Floræ, biRte and RABIT also rank *PRDM1* highly (ranks 3, 4 and 5). This TF silences stem cell-related genes and inhibits the proliferation of colon tumors [Kang et al., 2016; Liu et al., 2018; Zhu et al., 2017]. Even though the number of loops in the human TF-gene network is larger than in the E.coli network, there are still only 60 loops considered in our analysis of the COAD data set, and thus the results of Floræ are very similar compared to biRte. The three TFs in the top 10 of Floræ, that are not already present in biRte's top 10 (*MYC*, *TP53* and *SP3*), are all TFs within a loop and are rather generally altered TFs in cancer [Dang, 2012; Li and Davie, 2010; Parikh et al., 2014; Petitjean et al., 2007]. These findings suggest that Floræ identifies relevant TF in cancer patient data. However, the lack of a gold standard persists as a general problem in the evaluation of real biological data sets and therefore a quantitative assessment of the results remains difficult.

RACER	RABIT	biRte	Floræ
HOXA5	*MYC*	*AHR**	*AHR**
SP4	*KLF5*	*NR1I3**	*NR1I3**
MECOM	*CDX2*	*KLF5*	*PRDM1*
MLXIPL	*NRF1*	*PRDM1*	*CDX1*
CDX2	*PRDM1*	*CDX1*	***PHOX2B***
NRF1	*NFYA*	***PHOX2B***	*MYC*
MYC.MAX.ZBTB17	*NFKB1*	*ESRRA*	*TP53*
PHOX2B	***PHOX2B***	*HOXA5*	*SP3*
HOXA10	*RARG*	*TCF7L2*	*KLF5*
MYC	*PITX2*	*SOX2*	*ESRRA*

Table 5.4.: HGNC Symbols of the top 10 regulators found by RACER, RABIT, biRte and Floræ for the COAD data (165 samples). TFs with equal activity values are marked with asterisk. TFs found by several method's top 10 are marked in bold (when found by all four methods), blue (found by three methods) or red (found by two methods). Underlined TFs are part of at least one loop of length two or length four in the text mining network.

We apply Floræ to the knockout and knockdown data sets from human and E. coli cells described in Chapter 4. The results are given in Table 5.5, together with the number of loops of length two and four, that contain the KO or KD TF. We do not observe any relevant changes when comparing the results of biRte and Floræ. All TFs that were already ranked in the top 5% or top 5-10% of all ranked TFs by biRte are also retrieved by Floræ. In *FOXM1*, *STAT3* (SNB19 cells), *Fnr*, *OxyR* (aerobic condition) and *SoxS* (anaerobic condition), the ranks provided by Floræ are smaller than those from biRte. The maximal shift appears in *SoxS* from rank 14 to 11. In two cases, the combined *ArcA* & *Fnr* KD and *SoxS* (aerobic condition), Floræ ranks the KD TF on rank 2, instead of rank 1 like biRte. However, Floræ assigns on average smaller ranks to the searched TFs, but is not able to identify more KD or KO TFs as biRte.

Organism	Experiment	KO/ KD TF	Cell line/ condition	# of loops $l = 2$	# of loops $l = 4$	rank biRte	rank Floræ
Human	GSE45838	***BCL6***	OCI-Ly7	1	0	266	259
			Pfeiffer			163	142
	GSE17172	***FOXM1***	ST486	1	5	9	7
		MYB	ST486	0	13	112	181
	GSE19114	***bHLH-B2***	SNB19	1	0	186	158
		FOSL2	SNB19	0	0	355	340
		RUNX1	SNB19	0	0	8	8
		C/EBPβ	SNB19	0	0	-	137
			BTICs			328	331
		STAT3	SNB19	0	0	4	3
			BTICs			209	195
		C/EBPβ & STAT3	SNB19	0	0	-/ 31	209/ 31
			BTICs			402/188	387/180
E. coli	GSE1121	***AppY***	aerobic	0	0	119	117
			anaerobic			15	15
		ArcA	aerobic	1	0	198	191
			anaerobic			1	1
		ArcA & Fnr	aerobic	2	1	6 / 7	6 / 7
			anaerobic			1 /148	2 /158
		Fnr	aerobic	1	1	9	7
			anaerobic			192	157
		OxyR	aerobic	0	0	7	6
			anaerobic			6	6
		SoxS	aerobic	1	1	1	2
			anaerobic			14	11

Table 5.5.: Number of loops in which a KO or KD TF is part of (for loops of length two or four) and ranks of knocked down TFs per method and data set. Ranks in the top 5% of all ranked TFs are marked in green and ranks in the top 5–10% in light green. Two ranks in one table cell refer to a combined knockdown of two TFs and are given in the order of the TFs at the beginning of the table row. A dash is shown when a TF was not ranked by a method

5.3. Discussion

Feedback loops, despite their known importance for gene regulation [Alon, 2007; Komili and Silver, 2008], have rarely been considered in methods for the estimation of regulatory activity yet. Only a recently published study by [Kel et al., 2019] focuses on finding positive feedback loops in signal transduction pathways from expression data in colorectal cancer. They identified and experimentally validated six potential biomarkers of DNA methylation leading to rapid tumor development, hinting to the need for methods incorporating self-regulation. We developed Floræ (Feedback loops in regulatory activity estimation) as a new approach for estimating the activity of transcription factors with a particular focus on the consideration of feedback loops in the underlying gene regulatory network. Floræ is constructed modularly to facilitate the adaptation to different use cases. We first analyze the simulated effectiveness of knockout and knockdown by investigating the protein concentration of the according KO and KD TFs given by GNW. As expected, the KO of a TF clearly shows the highest percentage change of protein concentration compared to the other TFs, whereas in the KD data sets, other TFs than the KD TF have sometimes equal proportional changes. Using the the artificial networks and the simulated expression data, we show that Floræ ranks on average over all networks 8 out of 10 KO TFs (and 7 out of 10 KD TFs) on the first or second position, yielding better results compared to the three other methods. When removing the feedback edges in the network, Floræ yields comparable results to biRte. We further observe that the number of samples has no crucial impact on the estimated TF activity ranks. Randomization of a certain amount of network edges results in a higher variability of the TF activity estimations, but Floræ, biRte and RABIT still yield acceptable results when changing 10% of the interactions. Here, Floræ is the best method and ranks 7 of the 10 KO TFs on rank 1 or 2 (median rank). When applied to real biological experiments (data see Chapter 4), the results from Floræ are close to those from biRte, and only marginal improvements can be detected for the knockdown data from human and E. coli cell lines.

5.3.1. Method

The EM algorithm is widely applied to incomplete data problems in systems biology, where finding an analytical solution is impossible or would very complex or time consuming. Examples of the application of the EM algorithm include the detection of TF co-activations [Luo and Wei, 2018], the discovery of mutational patterns in cancer [Tan and Zhou, 2018], the modeling of stochastic microtubule signals [Menon et al., 2018] and metabolic flux determination [Boghigian et al., 2010]. The EM algorithm is an efficient and extensible method with good convergence properties, which have been theoretically proven [Chrétien and Hero, 2008; Dempster et al., 1977; Wu, 1983]. For example, it was shown that for Gaussian mixture models with a suitably large mean separation, even a relatively poor initialization suffices for the EM algorithm to converge to a near-global optimal solution [Balakrishnan et al., 2017].

An approach to detect gene deregulation from expression data using EM, which is quite close to the idea of Floræ to estimate TF activities, was published by [Picchetti et al., 2015]. However, their method explicitly excludes the presence of cycles in the network. We here propose a possibility to not only include but specifically use the information about feedback loops to improve the accuracy of TF activity estimation. We consider a mixture of Gaussian distributions, but the normality assumption might be violated in specific genes or TFs, especially at low sample sizes. Particularly in the case of a knockdown or knockout of a TF, the corresponding TF activities and gene expression values could be distributed differently. It might be useful to extend the method to detect such cases and change the mixture model accordingly, for example by applying a statistical test of normality or any other relevant distribution on the gene expression values and use an appropriate mixture model, like a mixture of gamma distributions.

We use biRte [Fröhlich, 2015] to provide initial TF activity estimations for Floræ and final estimations for TFs not included in a loop. We chose this method due to its relatively good performance in our previous analyses (see Chapter 4) and its comprehensive model. Of course, it would be possible to use other methods for this purpose and it would be easy to incorporate this change in the implementation of Floræ; only the computed values of TF activity should lie in the same range to assure a meaningful ranking. Currently, Floræ only examines loops of even length, considering just directed TF-gene or gene-TF network edges. It would be possible to extend Floræ to include cycles of uneven length $l \geq 3$ including e.g. TF-TF interactions, for example by inserting a dummy gene expression node in between, by combining both TFs in one node or by adding an intermediate EM run for the remaining edge. Note that we do not use the TF's mRNA levels as initialization for TF activity or as convergence criterion during the EM runs, since they are not a reliable proxy for TF activity due to effects of co-factors or post-transcriptional modifications [Brent, 2016].

The scoring method we use to compute the final activity values obviously has a significant influence on the results. We are interested in scoring TFs highly, that show a high activity in all samples or which have a high differential activity between two sample groups. However, the scoring could as well focus only on the TFs with differential activity, when applied with that scope. We currently do not check whether the sample group assignments at the end of the EM runs correspond to the original allocation in case and control samples. The accordance with the true assignment could also be used to designate a confidence score. Further, in the current version of Floræ, we do not compare the behavior of the predicted gene expression values to the measured ones during the EM runs. The distance between measured and predicted gene expression values could be used to improve the scoring by assessing the reliability and quality of the activities provided by Floræ compared to the ones given by the initialization method.

5.3.2. Data sets

We base our evaluation of Floræ mainly on synthetic data. We created five small regulatory networks, each including five feedback loops, and used the tool GeneNetWeaver

(GNW) [Schaffter et al., 2011] to simulate according gene expression data for wild-type, knockdown and knockout samples. Other data generators exist, like Netsim [Di Camillo et al., 2009] or SynTren [Van den Bulcke et al., 2006]. It would be interesting to compare the results from Floræ when using a different kinetic model in GNW or a different simulator as starting point for the evaluation. Another interesting objective would be to analyze the influence of the signal-to-noise ratio in the expression data on the results. In general, the use of synthetic data allows us to have a complete insight in the underlying model and gives us the control over all parameters. Experiments based on simulated networks and expression data are essential to assess the performance of inference methods [Marbach et al., 2009]. However, additionally to the already abstract formulation of mathematical models and the representation as networks in methods estimating regulatory activity, the simulation of synthetic data is a further simplification of biological reality.

We also use real biological data to evaluate Floræ. The results from Floræ are close to those from biRte, especially for the KO and KD data from human and E. coli, and only marginal improvements can be detected. One reason is that the KD and KO TFs are only partly included in the loops in the networks, and that the network from E. coli only includes a few loops that can be analyzed at all. A more detailed topological analysis of the underlying regulatory networks might help understand the limitations of Floræ in these cases. Further, when comparing the results of different methods by searching the literature for commonly found TFs, we inherently can only find already existing knowledge, restricting the explanatory power of our analyses. An additional problem is the general lack of large gene expression data sets with measurements of multiple samples per group, as already pointed out in Chapter 4. The data limitation will probably be removed over the next few years, now that the CRISPR/Cas9 technology has made deleting TFs in mammalian systems much easier [Sternberg and Doudna, 2015].

5.3.3. Networks

We use artificial networks purposefully designed to evaluate the performance of Floræ. All five networks include 10 TFs, 24 genes, 37 or 38 directed TF-gene and gene-TF interactions and five feedback loops of length 2 and/ or 4. We use networks of this small size and limit the feedback loop length to four to be able to interpret the effects of feedback. We use five different networks and the according simulated expression data as input to the methods estimating TF activity. The networks differ in the position of the feedback loops, have partly overlapping loops and loops of different length. However, the structures of the artificial networks are still similar to each other, and it would be interesting to evaluate the methods on networks with different size, different density or different connectedness. Since in reality the underlying regulatory network is, at least in mammals, not known completely, we randomize up to 50% of the edges of network A to study the influence of incomplete and partly incorrect network interactions. Note that our randomization procedure is not degree-preserving, leading to structural different networks and implying potential secondary effects. The randomization rate seems high,

but reflects a realistic scenario, as for example in human, the currently existing regulatory networks contain only a part of the estimated 1,500 TFs in the mammalian genome [Vaquerizas et al., 2009]. The incompleteness of the underlying networks might also be a reason for the limited performance of Floræ when using real biological data as input. Also [Klinger and Blüthgen, 2018] shows that good results of modular response analysis, a technique for finding connections between network modules, are restricted to the use of small sparse networks. Therefore, larger TF-gene networks could be used as input for the methods, like Reg-Network [Liu et al., 2015] or the one assembled from [Garcia-Alonso et al., 2019]. Further, the quality of existing networks might be improved by the integration of knowledge about the expression of enhancer RNAs and data on three-dimensional chromosome conformation [Kang et al., 2016]. Floræ currently focuses on TF-gene networks, as TFs influence gene regulation mainly. However, Floræ could also be applied to miRNA-gene interactions.

5.4. Conclusion

We presented a method for estimating the activity of transcription factors considering the influence of feedback loops in gene regulatory networks. To our knowledge, Floræ is the first method handling feedback loops to compute activity estimations, providing an important extension of previous methods. The results of our analyses show that Floræ can improve the identification of knockout and knockdown TFs in synthetic data sets compared to three other state-of-the-art methods. Additionally, we studied the effects of sample size and network randomization, where Floræ behaves similarly to the other methods.

Currently, Floræ is built modularly: It is based on another method estimating initial regulatory activity and subsequently performs an optimization of activity estimations in each feedback loop. Floræ implements an Expectation Maximization (EM) procedure, applying a Gaussian mixture model to the activity values of each TF in a feedback loop and the related gene expression values. Iteratively, we estimate the means and proportions of each Gaussian distribution of the mixture and use them to score TF activity. Of course, the details of the treatment of the feedback loops could be enhanced in different manners, and we discussed several possible adaptations of Floræ to adjust the method to different situations or conditions.

As long as it is too expensive or unfeasible to measure TF activity at large scale by protein mass spectrometry and transcription rates for targets directly [Brent, 2016], methods for estimating regulatory activity are necessary and valuable tools to describe causes and mechanisms of regulatory (dys)functions. Reliable activity estimations support the identification of biomarkers and the development of new therapies.

6. Conclusion

6.1. Summary

In this thesis, we studied methods for the estimation of regulatory activity using mathematical optimization. We compared several of such methods qualitatively in detail, explaining their common features and pointing out their differences. Since their results in the original publications were not comparable, we conducted a quantitative evaluation using the same input data and evaluation metric. We showed that the combined results for cancer patient data were biologically meaningful, but highly heterogeneous. Unfortunately, the methods were not able to robustly detect knocked down transcription factors. To take into account the previously ignored influence of feedback loops, we presented a novel method considering the effects of self-regulation. We showed that our method was able to improve the identification of knockout and knockdown TFs in synthetic data sets.

In **Chapter 2**, we introduced relevant biological and technical concepts for this thesis. We described the mechanisms of gene expression as well as measurement techniques. The role of different processes relevant for gene regulation and the concept of gene regulatory networks were presented. We described different methods for gene regulatory network reconstruction and introduced the problem of activity inference.

Chapter 3 described five recent methods for estimating genome-wide gene regulatory activity. We compared the published methods in detail with respect to the input data sets, the mathematical model, the technique to derive optimized activity values, the output and the evaluation procedures. We highlighted the common ground of the methods by illustrating the similarities and differences to a basic mathematical framework for activity inference. In this chapter, we showed that the presented methods, despite their enormous simplification of the underlying biological processes, were able to detect strong signals, facilitating hypotheses formulation for further research and being useful for the identification of biomarkers for specific phenotypes.

Since the results of the methods presented in Chapter 3 were not directly comparable, we conducted the first quantitative comparative evaluation of activity inference methods. **Chapter 4** presented the publicly available data sets and background networks used for this comparison and showed that the resulting activity ranks from different methods were highly divergent when investigating (multi-) omics patient cancer data. The result overlaps were small, though biologically meaningful and in some cases statistically significant. We further assessed data with lower biological complexity and compared the methods based on knockdown data of transcription factors in cell lines. We showed that

the methods were only rarely able to identify knockdown transcription factors and investigated the influence of network size and topology on the results. We discussed several limitations of the methods and our evaluation and suspected that the simplistic model of cellular processes used even in the more complex methods, ignoring self-regulation and feedback loops, was at least partly responsible for the limited performance.

We therefore devised in **Chapter 5** a novel method, Floræ (Feedback loops in regulatory activity estimation), to specifically include the effects of feedback loops from the underlying regulatory network which were not considered in any activity estimation method before. Floræ is built modularly and is based on an expectation maximization algorithm. Using synthetic networks and expression data, we showed that Floræ improved the identification of knockout and knockdown transcription factors. In the application to real biological data, the results from Floræ were comparable to those of other methods and could not substantially enhance the uncovering of regulatory interactions. We further investigated the influence of sample size and network randomization on the results, indicating that the results from Floræ were stable even for small sample sizes and when rewiring ten percent of the network edges. We finally discussed several potential extensions of Floræ, including modeling assumptions, and the limitations of synthetic data.

6.2. Future Directions

In the course of our research, the growing insight into the functioning and the performance of activity inference methods led us to further questions, and thus the results of this thesis point to multiple future research directions. This section critically discusses our achievements and gives an overview of several aspects which could be investigated in the future to achieve the overall goal of uncovering regulatory interactions, elucidating disease mechanisms, determining biomarkers for prognosis and diagnosis and eventually identifying therapeutic targets.

6.2.1. Experimental Data

Obviously, the selection and quality of the chosen experiments and their according data sets affect the type and quality of the outcome of the methods [Wang and Huang, 2014]. Throughout the work on this thesis, we chose to analyze different data types, coming from different species, from various diseases, from several cells of origin and cell lines and from different contributors, to make our results and conclusions less dependent on the specificities of a single data set. However, we did not check the quality of the input data in detail but relied on the high standards of the public data repositories, like TCGA and GEO. The use of other data sets or other data types could have led to different results, but we are confident, that the overall picture of the (in)capabilities of activity inference methods would not have been changed.

Sample Size

While the number of available patient data from TCGA was quite large, the number of biological replicates for one condition in the knockdown experiments was small, as by default in experimental practice [Gauthier et al., 2018]. Here, a high biological variability in only a few samples could have impaired the performance of the activity estimation methods, and larger experiments could be advantageous. On the other hand, reliably inferring activity from a single sample would be a cost-effective and in regards to animal experiments an ethically tenable option to gain knowledge about regulatory processes. Single sample results would also be interesting for the screening of unknown cases in a prognostic setting and should be investigated further [Li et al., 2014].

Data Types and Multi-omics

Using multi-omics data as input to regulatory inference methods seems meaningful, as e.g. a specific disease is rarely caused by a single gene but is rather a product of the interplay of genetic variability, epigenetic modifications and post-transcriptional regulation mechanisms. Measuring data on these different levels and integrating this knowledge offers an unprecedented opportunity to study how genetic information is used to control complex biological processes and their interaction [Davidsen et al., 2016]. In this thesis, we did not consider the effects of chromatin remodeling via histone modifications describing DNA accessibility and ignored the influence of complex promoter structures, especially distal promoters. However, these mechanisms present interesting possibilities for further research where measured data or computational models, which are able to predict genome-wide DNA accessibility and enhancer activity in terms of local constellations of regulatory sites, could be integrated to improve the estimation of regulatory activity. [Balwierz et al., 2014]. In recent years, the ability to generate omics data grew and many data sets have been published [Pataskar and Tiwari, 2016]. We used multi-omics data from the TCGA database as input for the investigated methods, but they all differed in the specific integration possibilities and we could not draw a definite conclusion about which combination of data types would currently be the most beneficial for activity inference. Further, the integration of different omics data is generally challenging since no standard or universal experimental methods, statistical procedures or computational analyses tools for facilitating integrative research exist [Delgado and Gómez-Vela, 2019]. This envisaged synergy requires both biologists and computer scientists to work together and share not only data, but also expertise, knowledge and processes [Thomas and Jin, 2014]. In the future, the in silico reconstruction of whole living organisms and their environments is conceivable, as already achieved in a whole cell model of a human pathogen [Gauthier et al., 2018].

In this thesis, we only used data drawn from populations of cells rather than individual cells, possibly entailing averaging artifacts [Brent, 2016]. Therefore, it would be desirable to enable activity inference methods to use single-cell RNA-Seq data, which recently become available at low-cost and can be generated highly parallel [Macosko et al., 2015]. For example, data about chromatin accessibility from Hi-C measurements

could be used to analyze higher order chromatin structures and thereby gene regulation at the single-cell level. However, these data sets are generally very noisy and require a large number of replicates to reach conclusions [Pataskar and Tiwari, 2016]. The successful adaptation of activity inference methods to single cell analysis would enable new biological insights into cell differentiation, cell-to-cell variation and gene regulation, as well as the interdependence of these aspects. [Hebenstreit, 2012].

Time series

Time series data has been studied in detail in gene regulatory network reconstruction [Berestovsky and Nakhleh, 2013; Schulz et al., 2012], but only rarely in activity inference [Balwierz et al., 2014; Jargosch et al., 2016]. Although synthetic time series data is publically available, for example from the DREAM challenges [Marbach et al., 2012; Meyer et al., 2014], or could be simulated with e.g. GeneNetWeaver [Schaffter et al., 2011], we did not find any suitable data of this type derived from real biological experiments. Further, only ISMARA could have been applied directly to time series data, whereas the other investigated methods in this thesis cannot use this type of data. However, the modeling of network dynamics could help detecting changes in regulatory activity over time and thereby improve the understanding of fundamental processes occurring in living organisms.

Overall, we expect that the efficacy of activity inference methods will further increase, as new functional genomics technologies develop and approaches to model the interactions between different layers of biological organization advance.

6.2.2. Background Networks

As pointed out in Chapter 3, a necessary input to the models is the underlying regulatory network, and the results clearly depend on its quality. Their construction is difficult for various reasons: first of all, not all TFs and miRNAs are known, especially in human. The main technique for the determination of TF binding is ChIP, which is known for its high false positive rate and generally not available for many TFs. Furthermore, many of the binding events identified by ChIP are nonspecific and many of those that are specific are nonfunctional [Lenstra and Holstege, 2012]. Computational prediction of transcription factor binding is also debatable [Jayaram et al., 2016]. For the evaluation of human data, we used a text mining based network published by [Thomas et al., 2015] and complemented it with ChIP data from TRANSFAC [Wingender et al., 1996]. At the time we performed our analyses, this was one of the largest publically available network, containing around 430 human TFs. Recently, [Garcia-Alonso et al., 2019] published a human gene regulatory network containing 1541 TFs, together with confidence scores for each of the around one million interactions, which seems to be the largest collection of human TF-target interactions. It would be interesting to use this network as input to the investigated activity inference methods and compare the results to our investigation, as well in regard to the computation time when using such a large network. Further,

the quality of existing networks might be improved by the integration of enhancer RNAs and data on three-dimensional chromosome conformation [Kang et al., 2016].

The graph representation of regulatory interactions is intuitive, but cannot account for TF complex formation and ignores also the required presence of co-factors for transcriptional regulation. The consideration of pairs or a higher number of combinations of TFs and the inclusion of information on temporal and spatial synchronization of TFs would require the integration of hypergraphs in the methods. Further, it is currently assumed, that a TF acts either mainly as activator or as repressor in activity inference models and gets assigned a single activity value, whereas it is clear that some TFs can have different effects on distinct targets [Balwierz et al., 2014]. Allowing for hypergraph edges and such a dual function could improve the results.

We conducted some experiments to assess the influence of the network topology on the results, for example by using smaller networks and de-novo inferred networks by ARACNE as input to the methods when studying the knockdown experiments from GEO (see Chapter 4). The reduction of the network size or the use of ARACNE's inferred networks did not improve the results. However, the effects of network incompleteness, error rates or the robustness to changes in the network could be evaluated further. In Chapter 5, we randomized 10% and 50% of the network edges and found that the variance of the results of all investigated methods increased with higher randomization rates. However, 50% alterations in the network does not seem an unrealistic scenario, since the existing knowledge of gene regulatory interactions is limited and partly condition specific. Therefore, activity inference methods would greatly benefit from a good quality of the background networks and more effort should be devoted to the assembly of high confidence regulatory networks.

6.2.3. Evaluation Procedure

For the estimation of regulatory activity, no gold standard, i.e. a data set on which the improvement of the accuracy of the results can be determined, is available to assess the performance of any given inference method [Brent, 2016]. We, inter alia, evaluated our results by searching highly ranked TFs in the existing biological literature. However, this strategy is flawed since we inherently can only find results which are already known. As often negative experimental results are not published at all [Fanelli, 2012], we might not know about negative results for TFs we consider to be biologically relevant. We also assessed the overlap of the results of different methods which was sometimes statistically significant. The implementation of an ensemble approach, scoring the results of different methods, could be a solution to increase the robustness of the results and to compensate strengths and weaknesses of all methods. We further used knockout and knockdown experiments to circumvent the problem of the absence of a gold standard data set. We expected that the methods would be able to identify the eliminated TF and that we could draw conclusions about the accuracy of the method. However, this was not the case, and we speculated that the knockdown affects only a small proportion of the whole

gene expression and the effect of deleting a TF on the expression levels of other genes seems to dissipate quickly in the network [Brent, 2016]. The general lack of publically available knockdown data sets including multiple samples per group might be removed over the next few years, now that the CRISPR/Cas9 technology has made deleting TFs in mammalian systems much easier [Sternberg and Doudna, 2015]. Another interesting evaluation approach, that we did not pursue, would be to use a clinical data set and show that the found features are meaningful predictors of outcome, for example by predicting the survival in a Kaplan-Meier analysis or any other clinical metric, like a tumor grade. Additionally, we did not evaluate our results using the i-score method [Berchtold et al., 2016], which systematically evaluates the inferred activity changes and thereby can rate the results of different methods. However, [Berchtold et al., 2016] stated that published methods yielded large, i.e. unfavorable, i-scores, meaning that the expression of many genes could not be explained by the set of activity changes of TFs. This negative result confirms our observations. Given that we could not identify a best method and that we are still far from developing realistic quantitative models of genome-wide gene regulatory dynamics in higher organisms, the most constructive contribution that computational approaches can currently provide is to develop models that help guide experimental efforts [Balwierz et al., 2014].

6.2.4. Extensions for Floræ

We proposed Floræ as a method considering the effect of feedback loops in regulatory activity estimation. Currently, Floræ computes only activity values for TFs in a loop, but should be extended to calculate values for all regulators, which we did not implement yet due to practical limitations of computation time. Also, it would be straightforward to adapt Floræ to, at present ignored, cycles with more than four edges and cycles of uneven length representing e.g. TF-TF interactions, for example by inserting a dummy gene expression node in between, by combining both TFs in one node or by adding an intermediate EM run for the remaining edge. The assumption of a mixture of Gaussian distributions might be violated in many cases, especially when a TF is knocked down or knocked out, and the performance of Floræ could be improved by detecting such cases and changing the model accordingly. In general, the distance between measured and predicted gene expression values could be used to improve the scoring by assessing the reliability and quality of the activities provided by Floræ compared to the ones given by the initialization method, which could itself be varied to adapt the method to different use cases. The scoring heuristic has a crucial influence on the results and should be examined carefully. Currently, we equally score TFs with high activity in all samples and TFs with high differential activity highly, but these two cases could be rated separately as well. Generally, Floræ only uses mRNA data as input and should be updated to include other omics-data as well to reflect knowledge of the regulation of biological processes in a broader spectrum and on different levels. Further, the inclusion of information from time series data could enable Floræ to detect changes in TF activity over time and thereby improve the accuracy of the results.

6.2.5. Perspectives

Currently, TF activity can not be experimentally measured by any high-throughput method. However, the comparison of (inferred) TF activities with TF mRNA levels could be used to generate hypotheses about post-transcriptional regulation and might be linked to upstream signaling pathways, leading to the investigation of pathway activity by using knowledge of pathways and how they affect TFs [Clarke et al., 2018]. It remains to investigate to what extent the inference of TF activity levels enables the estimation of the activity of specific TF activity modifiers, such as kinases and phosphatases [Brent, 2016]. Today, pathway activities are often inferred by the transcription levels of their members, ignoring the hard-to-measure post-transcriptional and post-translational regulation [Garcia-Alonso et al., 2019]. Measurements approximating total TF activity using quantitative mass spectrometry could monitor protein levels of hundreds of TFs in a single sample and could be also used to quantify post-translational modifications, but their robustness and scalability is unclear [Brent, 2016].

Although gene regulatory network models are certainly a powerful tool for the understanding of biological systems [Delgado and Gómez-Vela, 2019], further research is necessary to contribute to the establishment of guidelines for the quantification of human TF activities and to enable the characterization of complex relations between biological entities.

A. Appendix

A.1. Related TFs

The table shows TFs related to the knocked down TF.

Human				
Knockdown TF	**TFs directly connected in regulatory network**	**Aliases (GeneCards)**	**TFs connected via pathway (SignaLink 2.0)**	**Interactions with other TFs (TcoF DB)**
BCL6	*BCL2L1, CCND2, FCER2, TP53, FOXO4, SPI1*	*ZBTB27, ZNF51, BCL5, LAZ3, BCL6A*	-	*BCL11A, BCL6B, CREBBP, IRF4, JUN, JUNB, JUND, MTA3, NCOR1, NCOR2, PATZ1, SPI1, TP53, TWIST1, ZBTB16, ZBTB7A, ZBTB7B*
FOXM1	*ESR1, TP53, VEGFA, MYC*	*FKHL16, HFH11, MPP2, MPHOSPH2, TRIDENT, FOXM1B, HNF-3, INS-1, MPP-2, PIG29, WIN*	-	*SMAD3, SP1, ZBTB3*
MYB	*ETS1, HOXA9, IRF1, JUN, JUND, ADA, CDK1, COL1A2, GATA3, GSTP1, IGFBP5, KIT, MYC, NR3C1, PAX5, PAX6, PRTN3, SIM2, SLC34A2, SNAI2, SP3, SPP1, SRSF2*	*Cmyb, EFG*	*CCNA1, CCNB1, NR3C1*	*CEBPE, CREBBP, HLF, MAF, NCOR1, SKI, SMARCA2, SP100*
BHLHB2	*ARNT, BHLHE41, ID1, MLH1, PER2, HIF1A, TP53, TP63, TP73, ARNT.HIF1A, ARNTL.CLOCK*	*STRA13, SHARP2, DEC1, BHLHE40, HLHB2*	*APC2, BRCA1, PIM1*	*ENO1, HIVEP1, MYOD1, NOC4L, SOX15, TCF3, ZHX1*
FOSL2	*BRCA1, FOSL1*	*FRA2*	-	*ATF2, ATF3, ATF7, CREB5, DDIT3, JUN, JUNB, JUND, MAFB*
RUNX1	*BCL2, CHI3L1, CLC, CSF1R, CSF2, FOXP3, GATA2, GPR132, GRAP2, IL19, IMPDH2, LAT, LGALS3, NCAM1, NFE2, PLAU, PRKCB, RUNX3*	*CBFA2, AML1, PEBP2αB, AMLCR1, EVI-1*	*ADRA1A, AGTR1, AHI1, ANXA1, ARHGAP39, B4E345, CDK6, CREBBP, EPHA3, GFRA1, GNAQ, IL21, ITGA1, LCP2, MAP3K8, NEK7, PDE3B, PPM1A, PPP2R2A, PRKCA, PRKCB, PRKCW, PTPRK, SEC31B, SH3BP5, SPAG16, SYK,*	*CBFB, CEBPB, DNMT1, ELF1, ELF2, ELF4, FOS, FOXP3, JUN, KAT6A, MYOD1, NCOR1, PAX5, SPI1, VDR*

			TLE1, TRIB1, WDR37, WDR70	
CEBPB	*ABCB1, BCL2A1, BCL2L1, CCL3, CCL4, CCL5, CCR5, CDKN1A, CYP19A1, CYP27B1, DDIT4, DUSP1, F7, FGFBP1, FGFR2, GFER, GLS2, HP, HSD17B8, HSPH1, IL1B, IL1RN, IL5, IL6, INSR, IRF9, LCN2, LDLR, MEFV, MMP1, PLAC1, PRLR, RUNX2, SAA1, SAA2, SLC19A1, SOX6, SPINK1, TNF, TNFAIP6, TNFRSF10, TOP1, TRAF3IP2,YWHAE, CFTR, GPX4, JUN, MMP2, TP63*	*TCF5, IL6DBP, NF-IL6, TCF-5, LAP, LIP*	*APC2, ARMC6, AURKA1, B7Z708, BRCA1, CCNK, CD7, CDC42BPA, CDC7, CLTC, CSTF1, CXCR4, DUSP1, IFT122, ITGB7, NEK3, NEK6, PARD6A, PCNA, PCSK6, PDIK1L, PIM1, POC1A, TRIB1, TRIB3, UBE2F*	*AR, ATF3, ATF4, BATF, BATF3, CEBPA, CEBPD, CEBPG, CREB1, CREBBP, DDIT3, EGR1, ESR1, FOXO1, HMGA1, HSF1, KLF5, NCOR2, NFKB1, NR3C1, PPARG, RARB, RELA, RUNX1, SMAD3, SMAD4, SMARCA2, SPI1, SRF*
STAT3	*HOXA1, AKT1, BIRC5, CCR5, CD274, CEBPD, CRP, CYP19A1, FAAH, FGG, FOS, HGF, IL10, IRF1, LBP, MMP1, MMP7, MUC1, MYD88, NOS3, REG1A, ROR1, SREBF1, TP53, TP63, VEGFA, VIM*	*APRF, ADMIO1, ADMIO, HIES*	*BIRC5, ESR1, ETS1, FOXA1, IL10, INSM1, IRF1, JM2, NHLH1, NOS3, NR2F1, PBX1, PLAG1, PRO2286, STAT1, TCF3, TEAD1, VDR*	*AR, ATF3, BATF3, CREBBP, GTF2I, HES1, HES5, HESX1, HIC1, HIVEP1, HOXC11, JUN, KLF15, MYOD1, NCOA1, NFKB1, NFKBIZ, NR3C1, NR4A1, PPARD, RELA, SMARCA4, STAT6, TWIST1, ZFPM2, ZNF281, ZNF557, ZNF829*
E. coli				
Knockdown TF	**TFs directly connected in regulatory network**	**Aliases (EcoCyc)**	**TFs connected via pathway (EcoCyc)**	
AppY	*DpiA, H-NS*	*-*	*RpoS, ArcA, H-NS, DpiA*	
ArcA	*Fnr*	*sfrA, cpxC, dye, fexA, msp, seg*	*Fnr*	
Fnr	*ArcA, Fur, IHF*	*nirA, nirR, ossA, oxrA*	*IHF, Fur, ArcA, SoxS*	
OxyR	*CRP*	*momR, mor*	*CRP*	
SoxS	*SoxR, AcrR, Fnr, Fur*	*-*	*SoxR, MgrR, AcrR, Fnr, Fur*	

A.2. Differential expression of KD TFs

The table indicates p-values for differential expression of KD TFs in human and E. coli.

Organism	Experiment	**TF knockdown**	Cell line/ condition	p-value
Human	GSE45838	***BCL6***	OCI-Ly7	0.001368
			Pfeiffer	0.000097
	GSE17172	***FOXM1***	ST486	0.000360
		MYB	ST486	0.000072
	GSE19114	***bHLH-B2***	SNB19	0.012660
		FOSL2	SNB19	0.009246
		RUNX1	SNB19	0.001226
		C/EBPβ	SNB19	3.94e-07
			BTICs	0.957500
		STAT3	SNB19	1.88e-08
			BTICs	3.42e-14
		C/EBPβ & STAT3	SNB19	3.84e-07 & 4.91e-08
			BTICs	0.05750 & 1.16e-12
E. coli	GSE1121	***AppY***	aerobic	0.027750
			anaerobic	0.010610
		ArcA	aerobic	0.014550
			anaerobic	0.002621
		ArcA & Fnr	aerobic	0.012500 & 0.002288
			anaerobic	0.002174 & 0.001488
		Fnr	aerobic	0.001773
			anaerobic	0.001363
		OxyR	aerobic	0.001363
			anaerobic	7.69e-06
		SoxS	aerobic	0.032830
			anaerobic	0.000152

A.3. Ranks of related TFs in KD data sets

The following table shows the ranks of KD TFs (bold) and related TFs, the total number of ranked TFs per method and a p-value indicating significance of the test whether the mean of the ranks of all related TFs is smaller than the average rank (total number of ranks divided by 2). Significant p-values are marked in yellow. Ranks of TFs in the top 5% of all ranked TFs are marked in dark green, ranks in the top 5-10% in green and ranks in the top 10-20% in light green. When a TF was not ranked, "-" is shown. Two ranks in one table cell refer to a combined KD of two TFs and are given in the order of the TFs at the beginning of the table row.

Human

Experiment GSE45838: knockdown of BCL6

Cell line	OCI−Ly7				Pfeiffer			
Method / TF	**biRte**	**ISMARA**	**RABIT**	**RACER**	**biRte**	**ISMARA**	**RABIT**	**RACER**
BCL6	**266**	**-**	**-**	**-**	**163**	**-**	**-**	**68**
TP53	397	138	-	-	-	235	-	-
FOXO4	49	374	-	33	1	97	-	-
SPI1	116	146	-	-	-	3	-	-
IRF4	36	-	17	-	117	-	-	-
JUN	-	321	39	83	386	116	-	-
JUNB	147	-	-	-	376	-	-	-
JUND	-	167	-	-	7	253	-	-
TWIST1	332	-	-	-	78	-	29	-
ZBTB16	377	392	-	11	179	485	-	42
ZBTB7A	282	-	-	-	397	-	-	-
ZBTB7B	353	-	-	10	262	-	-	-
total	404	500	58	88	405	500	53	143
p-value	0.770	0.512	0.471	0.368	0.453	0.181	-	0.212

Experiment GSE17172: knockdown of FOXM1

Cell line	ST486			
Method / TF	**biRte**	**ISMARA**	**RABIT**	**RACER**
FOXM1	**9**	**-**	**-**	**-**
ESR1	387	218	22	-
TP53	259	310	-	-
SMAD3	113	62	-	-
SP1	339	555	-	-
ZBTB3	-	558	-	-
total	398	602	63	4
p-value	0.617	0.649	-	-

Experiment GSE19114: knockdown of bHLH-B2, FOSL2, RUNX1, C/EBPβ, STAT3 and C/EBPβ & STAT3

Cell line	SNB19		
Method / TF	**biRte**	**RABIT**	**RACER**
bHLH-B2	**186**	**-**	**-**
ARNT	202	21	-
ID1	237	-	-
TP53	54	37	-
TP63	326	-	-
TP73	146	-	-
ARNT.HIF1A	263	-	-
BRCA1	6	6	-

ENO1	223	-	-			
HIVEP1	155	-	-			
TCF3	203	-	-			
total	402	42	0			
p-value	0.136	0.513	-			
Method / TF	**biRte**	**RABIT**	**RACER**			
FOSL2	**355**	-	-			
BRCA1	237	-	-			
FOSL1	276	54	-			
ATF2	321	-	-			
ATF3	384	11	-			
JUN	365	-	-			
JUNB	229	-	-			
JUND	261	-	-			
total	404	54	0			
p-value	0.999	0.58	-			
Method / TF	**biRte**	**RABIT**	**RACER**			
RUNX1	**8**	**37**	**-**			
FOXP3	213	-	-			
NFE2	271	6	-			
RUNX3	56	-	-			
CBFB	371	-	-			
CEBPB	6	17	-			
ELF1	251	-	-			
ELF2	9	1	-			
ELF4	257	-	-			
FOS	2	11	-			
MYOD1	346	-	-			
PAX5	255	-	-			
VDR	233	-	-			
total	401	49	0			
p-value	0.261	0.09	-			
Cell line	BTICs			SNB19		
Method / TF	**biRte**	**RABIT**	**RACER**	**biRte**	**RABIT**	**RACER**
CEBPB	**328**	**-**	**-**	**-**	**-**	**-**
RUNX2	-	-	-	290	-	-
JUN	385	-	-	394	-	-
TP63	56	-	-	-	-	-
BRCA1	262	-	-	166	-	-
AR	129	-	-	397	44	-
ATF3	238	6	-	268	-	-

ATF4	346	36	-	160	-	-
CEBPA	322	-	-	244	-	-
CEBPD	320	-	-	311	-	-
CREB1	269	-	-	-	-	-
EGR1	323	-	-	336	-	-
ESR1	370	44	-	371	-	-
FOXO1	90	-	-	365	-	-
HMGA1	200	-	-	347	-	-
HSF1	87	-	-	26	-	-
KLF5	291	-	-	28	38	-
NFKB1	118	-	-	-	-	-
NR3C1	-	-	-	143	41	-
PPARG	168	-	-	174	-	-
RARB	314	-	-	116	-	-
RELA	-	-	-	227	-	-
RUNX1	281	-	-	398	-	-
SMAD3	276	-	-	281	-	-
SMAD4	335	-	-	366	-	-
SMARCA2	306	-	-	291	-	-
SPI1	-	-	-	167	-	-
Max rank	397	61	14	404	49	0
p-value	0.989	0.444	-	0.982	0.995	-

Cell line	BTICs			SNB19		
Method / **TF**	**biRte**	**RABIT**	**RACER**	**biRte**	**RABIT**	**RACER**
STAT3	**209**	**-**	**-**	**4**	**29**	**-**
HOXA1	1	58	-	5	21	-
CEBPD	223	55	-	25	-	-
FOS	88	-	-	396	-	-
IRF1	213	-	-	-	-	-
MUC1	298	-	-	75	-	-
SREBF1	309	-	-	7	38	-
TP53	-	-	-	167	55	-
TP63	202	-	-	289	-	-
ESR1	-	-	-	326	-	-
ETS1	393	-	-	238	-	-
FOXA1	131	-	-	383	-	-
INSM1	343	-	-	289	-	-
NR2F1	346	-	-	377	-	-
STAT1	-	-	-	81	-	-
TCF3	157	-	-	35	-	-
TEAD1	321	51	-	176	-	-
VDR	170	-	-	403	-	-
AR	369	-	-	-	-	-

ATF3	123	56	-	257	-	-
GTF2I	54	-	-	72	-	-
HES1	332	-	-	363	-	-
HIVEP1	156	-	-	344	-	-
KLF15	178	-	-	288	-	-
MYOD1	162	-	-	400	-	-
NCOA1	383	-	-	225	-	-
NFKB1	356	-	-	323	18	-
NR4A1	82	17	-	351	-	-
PPARD	128	-	-	123	-	-
RELA	395	16	-	-	-	-
STAT6	103	-	-	102	-	-
TWIST1	11	-	-	2	23	-
ZNF281	184	-	-	31	33	-
total	405	60	14	403	59	0
p-value	0.700	0.902	-	0.562	0.617	-

Cell line	BTICs			SNB19		
Method / **TF**	**biRte**	**RABIT**	**RACER**	**biRte**	**RABIT**	**RACER**
STAT3	**188**	**-**	**-**	**31**	**-**	**-**
CEBPB	**402**	**-**	**-**	**-**	**-**	**-**
HOXA1	3	-	-	70	-	-
CEBPD	356	-	-	290	-	-
FOS	150	-	-	359	-	-
IRF1	71	47	-	-	-	-
MUC1	122	-	-	267	-	-
SREBF1	357	-	-	396	-	-
TP53	-	-	-	373	-	-
TP63	332	-	-	185	-	-
ESR1	232	-	-	-	-	-
ETS1	380	-	-	-	-	-
FOXA1	355	-	-	351	-	-
INSM1	320	-	-	82	-	-
NR2F1	370	-	-	52	-	-
STAT1	347	-	-	-	-	-
TCF3	336	-	-	138	-	-
TEAD1	322	-	-	273	-	-
VDR	40	60	-	312	-	-
AR	191	-	-	149	-	-
ATF3	353	8	-	-	-	-
GTF2I	57	-	-	369	-	-
HES1	186	-	-	216	-	-
HIVEP1	95	-	-	78	-	-
JUN	-	-	-	393	-	-

KLF15	304	-	-	183	-	-
MYOD1	70	-	-	204	-	-
NCOA1	133	-	-	334	-	-
NFKB1	45	-	-	253	-	-
NR3C1	403	42	-	320	-	-
NR4A1	124	16	-	200	-	-
PPARD	160	-	-	177	-	-
RELA	272	65	-	-	-	-
STAT6	299	-	-	197	-	-
TWIST1	8	2	-	3	3	-
ZNF281	178	-	-	311	-	-
RUNX2	309	-	-	368	-	-
BRCA1	17	34	-	300	-	-
ATF4	262	21	-	281	-	-
CREB1	391	-	-	387	-	-
EGR1	204	-	-	-	-	-
FOXO1	-	-	-	371	-	-
HMGA1	243	-	-	239	-	-
HSF1	288	-	-	54	-	-
KLF5	311	-	-	256	-	-
PPARG	406	-	-	106	50	-
RARB	146	-	-	102	-	-
RUNX1	201	-	-	-	-	-
SMAD3	22	41	-	-	28	-
SMAD4	20	37	-	-	-	-
SMARCA2	164	-	-	111	-	-
SPI1	69	-	-	394	-	-
SRF	405	-	-	373	-	-
total	410	71	14	400	51	0
p-value	0.797	0.40	-	0.972	0.539	-

E. coli

Experiment GSE1121: knockdown of AppY, ArcA, Arca & Fnr, Fnr, OxyR and Soxs

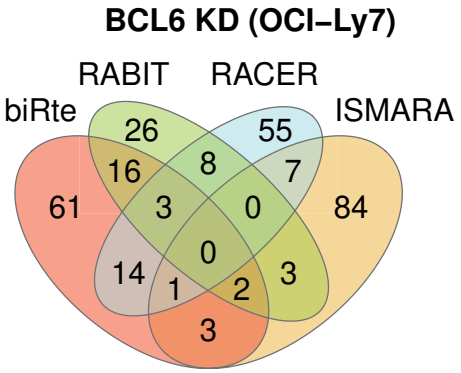
Condition	aerobic			anaerobic		
Method / **TF**	**biRte**	**RABIT**	**RACER**	**biRte**	**RABIT**	**RACER**
AppY	**119**	**-**	**73**	**15**	**1**	**71**
DpiA	148	-	-	24	-	-
H-NS	154	-	137	169	-	-
ArcA	14	21	43	1	24	64
total	199	48	152	198	43	121
p-value	0.603	-	0.604	0.159	0.289	0.852

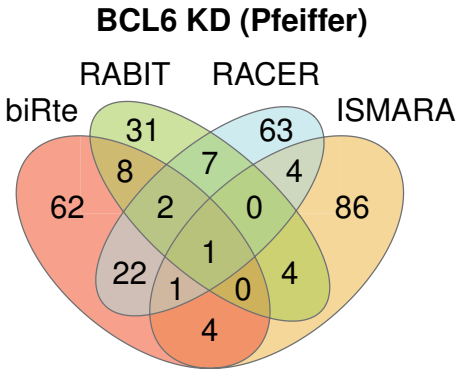
Condition	aerobic			anaerobic		
Method	**biRte**	**RABIT**	**RACER**	**biRte**	**RABIT**	**RACER**

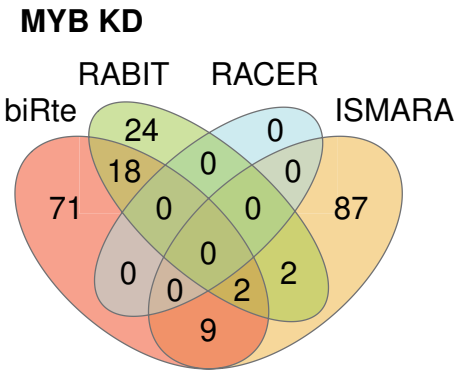
TF						
ArcA	**198**	-	**70**	**1**	**2**	**135**
Fnr	197	-	-	195	-	138
total	198	32	142	199	42	147
p-value	0.998	-	-	0.495	-	0.992
Condition	aerobic			anaerobic		
Method / **TF**	**biRte**	**RABIT**	**RACER**	**biRte**	**RABIT**	**RACER**
ArcA	**6**	**5**	**108**	**1**	**1**	**34**
Fnr	**7**	**6**	-	**148**	-	**104**
Fur	1	1	105	12	14	41
IHF	184	12	71	196	20	105
SoxS	14	24	103	198	-	101
total	199	29	133	198	45	115
p-value	0.091	0.144	0.98	0.597	0.097	0.853
Condition	aerobic			anaerobic		
Method / **TF**	**biRte**	**RABIT**	**RACER**	**biRte**	**RABIT**	**RACER**
Fnr	**9**	**10**	-	**192**	**43**	**127**
ArcA	7	2	99	1	6	37
Fur	1	1	118	46	45	51
IHF	186	18	59	196	31	111
SoxS	58	-	110	197	-	79
total	199	33	137	199	55	143
p-value	0.124	0.057	0.939	0.719	0.648	0.696
Condition	aerobic			anaerobic		
Method / **TF**	**biRte**	**RABIT**	**RACER**	**biRte**	**RABIT**	**RACER**
OxyR	**7**	**28**	**83**	**6**	**10**	**94**
CRP	179	-	107	5	-	-
total	197	34	135	199	35	121
p-value	0.48	-	0.869	0.002	-	-
Condition	aerobic			anaerobic		
Method / **TF**	**biRte**	**RABIT**	**RACER**	**biRte**	**RABIT**	**RACER**
SoxS	**1**	**10**	**95**	**14**	-	**92**
SoxR	11	-	-	1	-	-
AcrR	139	-	47	163	-	57
Fnr	190	15	-	53	5	101
Fur	1	-	52	182	12	37
total	199	40	146	199	45	119
p-value	0.241	0.102	0.32	0.337	0.078	0.763

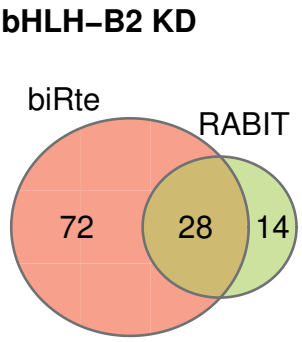
A.4. Overlap of the top 100 TFs in KD data sets

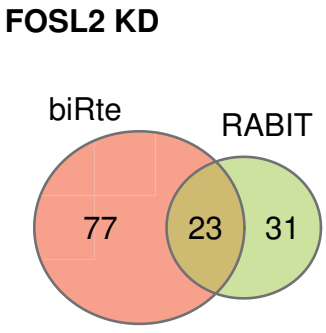
Venn diagrams show the number of overlapping TFs in the top 100 lists by estimating TF activity with different methods. For RABIT and RACER, the total number of ranked TFs was in some cases below 100 (see Table 4.4).

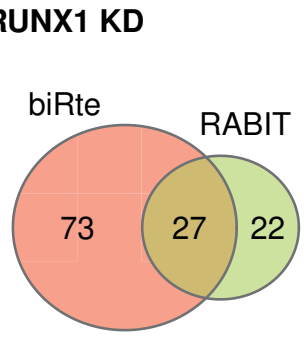




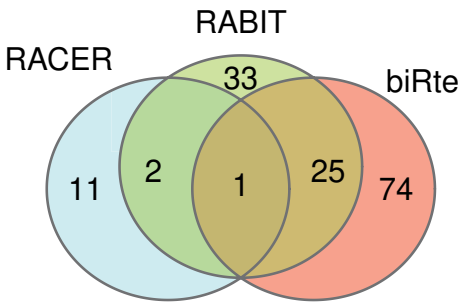




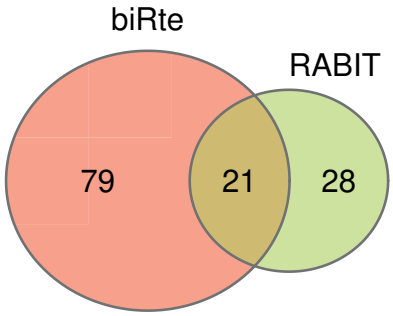




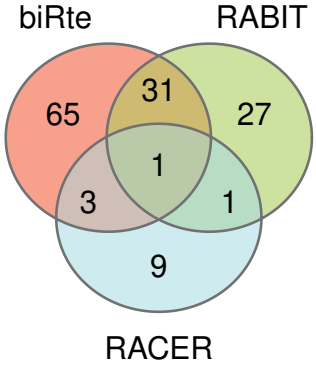
CEBPB KD (BTICs)



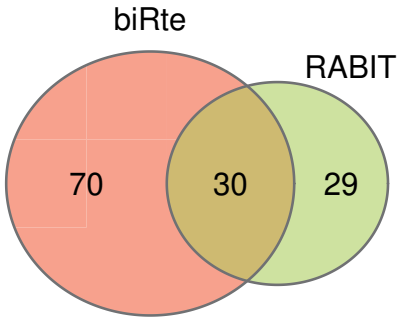
CEBPB KD (SNB19)



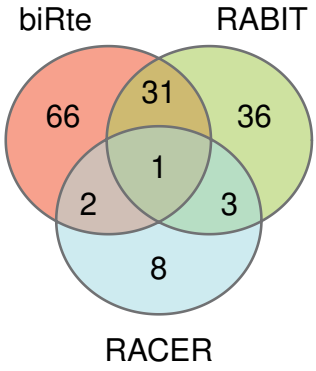
STAT3 KD (BTICs)



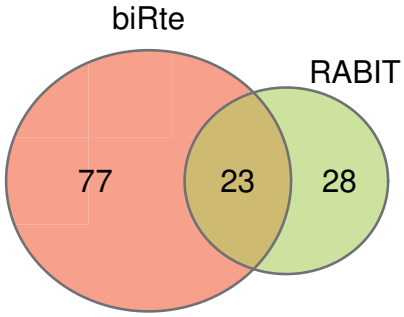
STAT3 KD (SNB19)

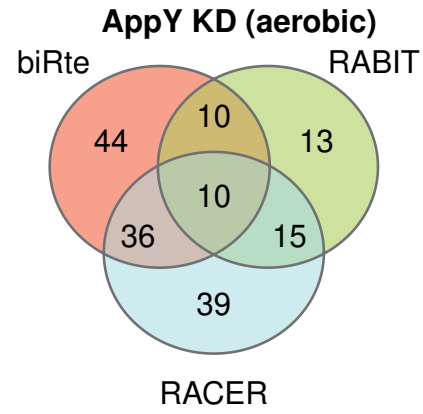


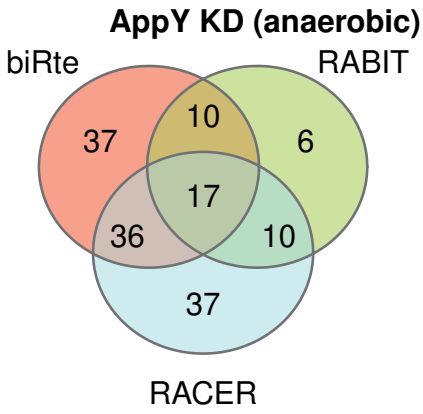
CEBPB & STAT3 KD (BTICs)

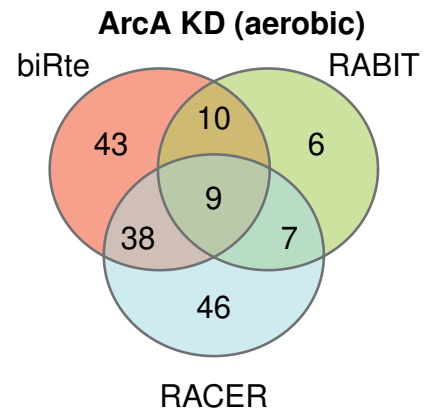


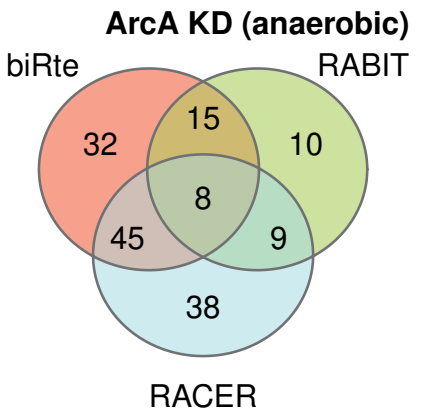
CEBPB & STAT3 KD (SNB19)

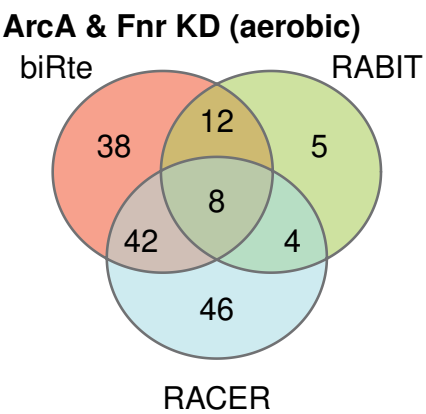


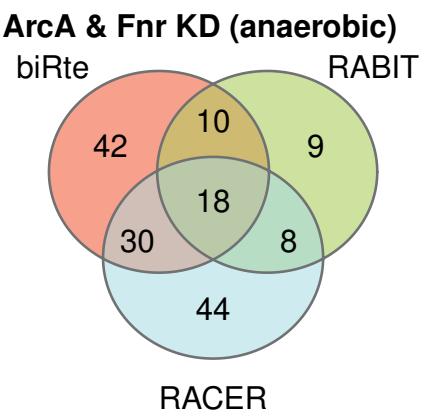

AppY KD (aerobic)
biRte
RABIT
44
10
13
10
36
15
39
RACER

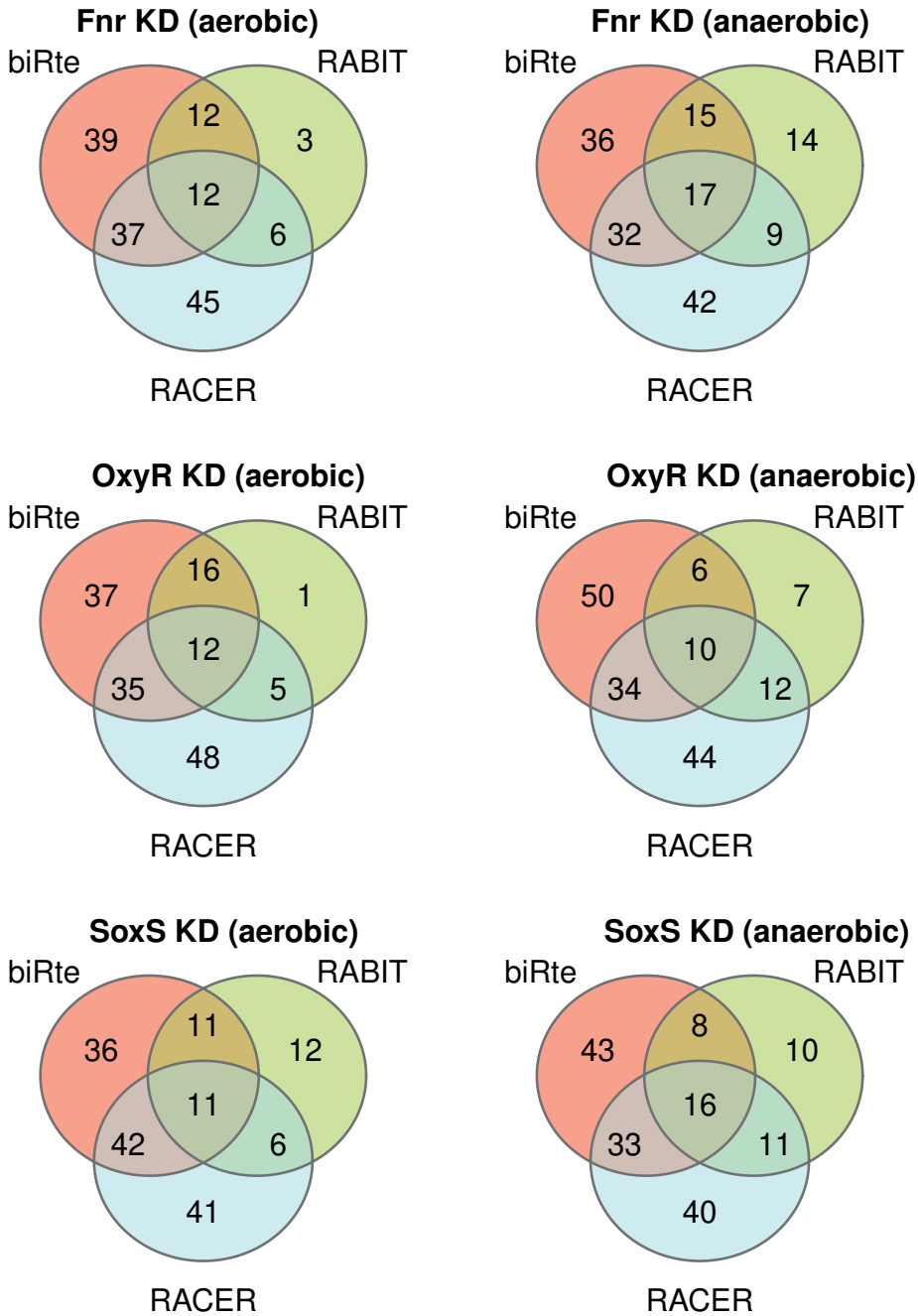

AppY KD (anaerobic)
biRte
RABIT
37
10
6
17
36
10
37
RACER


ArcA KD (aerobic)
biRte
RABIT
43
10
6
9
38
7
46
RACER


ArcA KD (anaerobic)
biRte
RABIT
32
15
10
8
45
9
38
RACER


ArcA & Fnr KD (aerobic)
biRte
RABIT
38
12
5
8
42
4
46
RACER


ArcA & Fnr KD (anaerobic)
biRte
RABIT
42
10
9
18
30
8
44
RACER


Fnr KD (aerobic)
biRte
RABIT
39
12
3
12
37
6
45
RACER
Fnr KD (anaerobic)
biRte
RABIT
36
15
14
17
32
9
42
RACER
OxyR KD (aerobic)
biRte
RABIT
37
16
1
12
35
5
48
RACER
OxyR KD (anaerobic)
biRte
RABIT
50
6
7
10
34
12
44
RACER
SoxS KD (aerobic)
biRte
RABIT
36
11
12
11
42
6
41
RACER
SoxS KD (anaerobic)
biRte
RABIT
43
8
10
16
33
11
40
RACER

A.5. Network properties

The table compares several network properties for the text mining network, the E.coli network and those networks inferred by ARACNE for the *FOXM1* KD (human) and the *AppY* KD (E.Coli).

Network	# of edges	Transitivity	# of nodes with hub score > 0.1	# of nodes with degree > 10
Text mining	2894	0.012	2	102
ARACNE (*FOXM1*)	59829	0.046	86	2283
E. coli	3954	0	9	92
ARACNE (*AppY*)	15540	0.028	20	379

A.6. Pseudocode of Floræ

Function `find_loops`

Algorithm 1: Find loops in the network

```
Function find_loops(network, max_length)
    Data: Network, maximum length of loops
    Result: List of loops with corresponding nodes
    - Load package igraph
    for l in seq(2,max_length, by=2) do
        all_cycles[l]=graph.get.subisomorphisms.vf2(network,
          graph.ring(l,directed=TRUE))
        # Keep only one version of each cycle
        cycles[l]=drop_permutations(all_cycles[l])
    end
    # Assemble all results in list
    loops=list(cycles[2:max_length])
    return loops
end
```

Function `apply_biRte`

Algorithm 2: Apply biRte to get initial TF activities

Function *apply_biRte(expression_data, network)*
  Data: mRNA expression data for case and control samples, network
  Result: Activities for each TF in each sample
  - Load package *birte*
  `# Data handling`
  n_cases= number of case samples
  n_controls=number of control samples
  - Match genes from network to probes from expression data
  - When multiple probes matched to one gene: Take probe with highest difference in p-value of t-test between case and control group
  - Reduce expression data to genes present in network
  `# BiRte analysis for each sample`
  TF_act=birteRun(dat.mRNA=*expression_data*, affiliations=*network*, nrep.mRNA=c(*n_cases*, *n_controls*),niter=10000, nburnin=10000, single.sample=TRUE)
  return *TF_act*
end

Function `EM_loops`

Algorithm 3: EM algorithm and final scoring

Function *EM_loops(expression_data, loops, TF_act, epsilon)*
- **Data:** mRNA expression data for all samples, loops of network, initial activity values for all samples and TFs from `apply_biRte`, threshold epsilon as convergence criterion
- **Result:** Sample specific TF activity scores for TFs in a loop, distribution parameters

`# Data handling`
- Same as in algorithm 2

n=number of samples

for *curr_loop* **in** *loops* **do**
  while change of $EM\_TF\_act > epsilon$ **do**
    for each TF - gene edge **in** *curr_loop* **do**
      Get names of $curr\_tf$ and *curr_gene*
      $expr$=$gene\_expression[curr\_gene]$
      `# Initialize` S, μ_s `and` p
      s=round($TF\_act[curr\_tf]$)
      mu_0=mean($expr[s == 0]$); mu_1=mean($expr[s == 1]$)
      p=mean($s == 0$)
      $sigma\_0$=sd($expr[s == 0]$); $sigma\_1$=sd($expr[s == 1]$)
      Q_new=1000; $niter = 0$
      while $(Q\_old - Q\_new) > epsilon$ **or** $niter < 1000$ **do**
        `# E step`
        $Q\_old = Q\_new$
        - Compute tau_0 and τ_1
        - Compute Q_new
        `# M step`
        - Compute p
        - Compute mu_0 and mu_1
        - Compute $sigma\_0$ and $sigma\_1$
        $niter$=$niter + 1$ `# Print warning when` $niter$=1000
      end
      `# Get activity estimations per sample`
      $EM\_TF\_act[curr\_tf]$=tau_0
    end
    for each gene - TF edge in *curr_loop* **do**
      - Repeat content of previous for-loop, but with reversed TF and gene assignments and EM_TF_act as input for $expr$, initialize s according to case and control samples.
      - As result, estimations for gene activity are obtained and stored in $expr$.
    end
  end
end
- When a TF occurs multiple times: Take maximal TF activity

return τ_0,mu,p

end

Function `scoring`

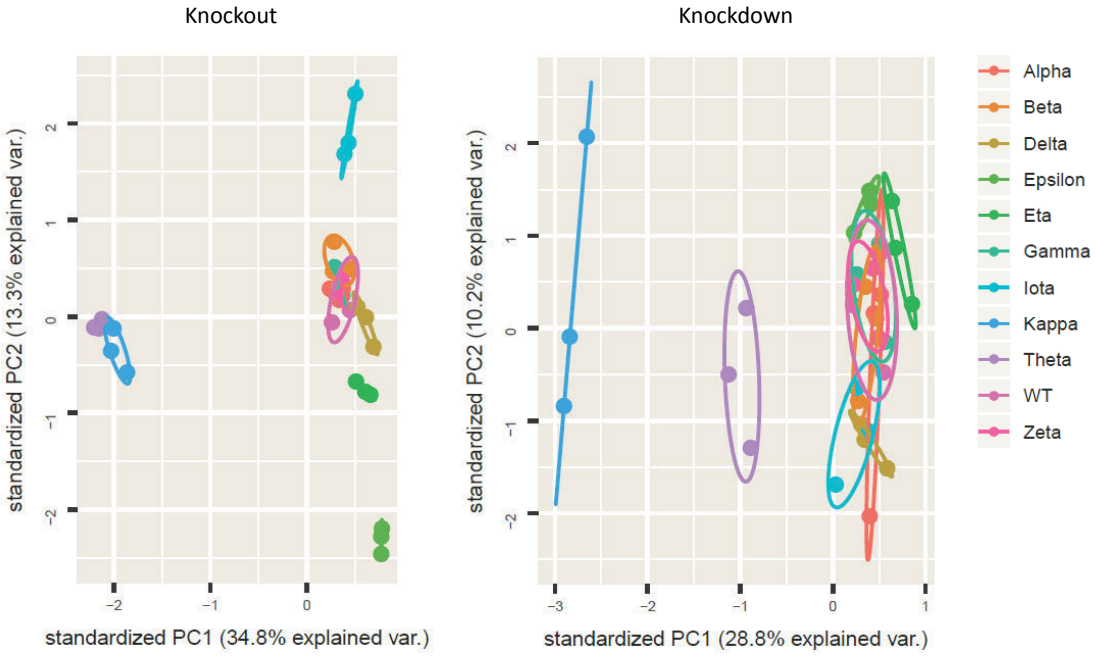
Algorithm 4: Score TF activity values and assemble results

Function *scoring(tau_0, mu, p, loops, expression_data, network)*
- **Data:** Estimated TF activity values per samples and distribution parameters from `EM_loops`, loops of network, mRNA expression data for case and control samples, network
- **Result:** TF activity values over all samples for all TFs
- Load package *birte*
- n=number of samples
- `# TFs in loops`
- **for** all TFs in loops (t) **do**
  - **if** *sum(tau_0[t]* $\geq$ *0.5)* $\geq$ *0.75*n* **then**
    - `# TF inactive`
    - TF_act[t]=mu_0[t]
  - **end**
  - **if** *sum(tau_0[t]* $\geq$ *0.5)* $\leq$ *0.25*n* **then**
    - `# TF active`
    - TF_act[t]=mu_1[t]
  - **end**
  - **if** *sum(tau_0[t]* $\geq$ *0.5)* $<$*0.75*n & >* *0.25*n* **then**
    - `# differential activity`
    - TF_act[t]=abs(mu_1[t]-mu_0[t])
  - **end**
- **end**
- **for** all TFs not in loop (t) **do**
  - Calculate TF activity using biRte method (TF_birte)
- **end**
- Assemble all TF_act and TF_birte values in one list: TF_act_final
- **return** TF_act_final

end

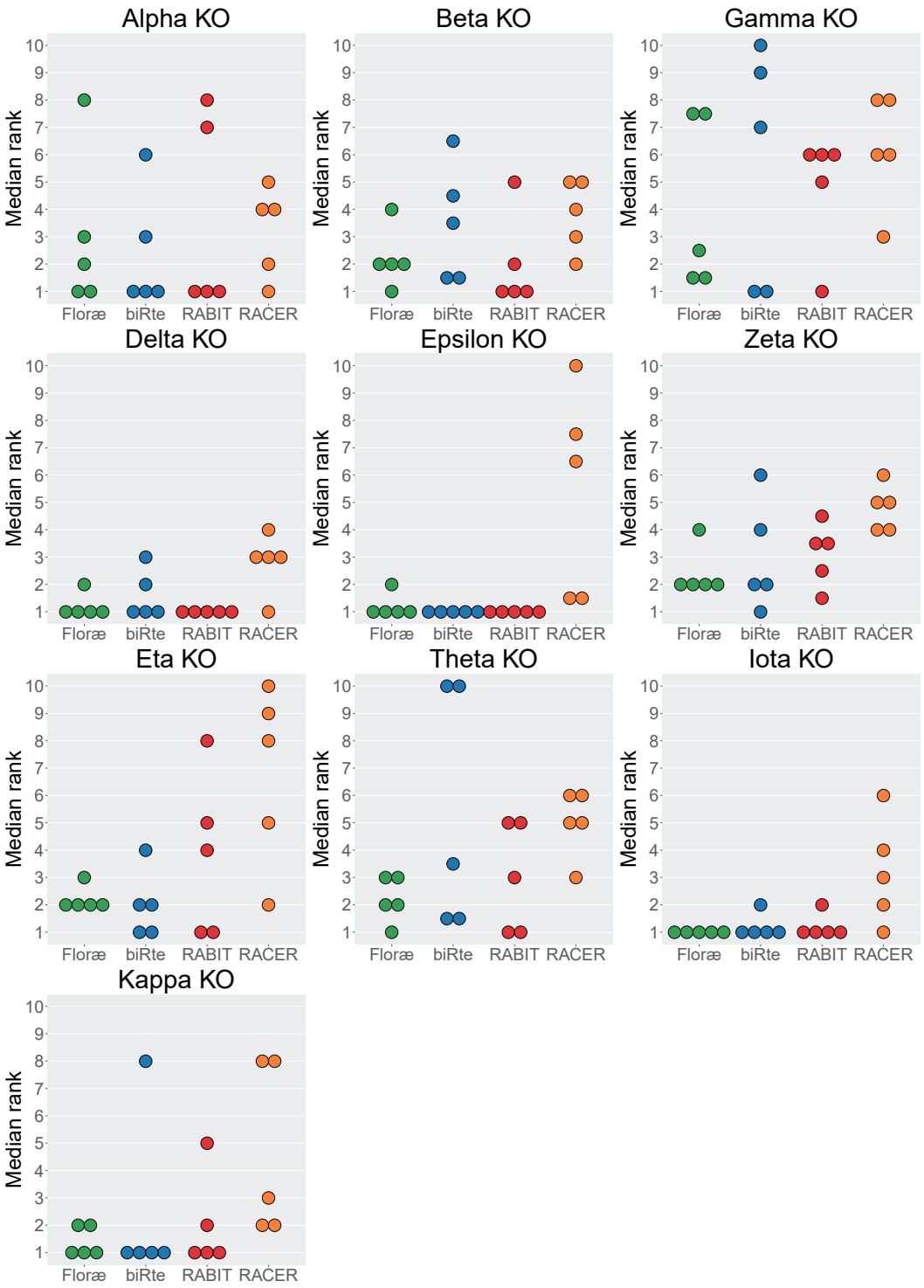
A.7. PCA of WT, KO and KD simulated expression data

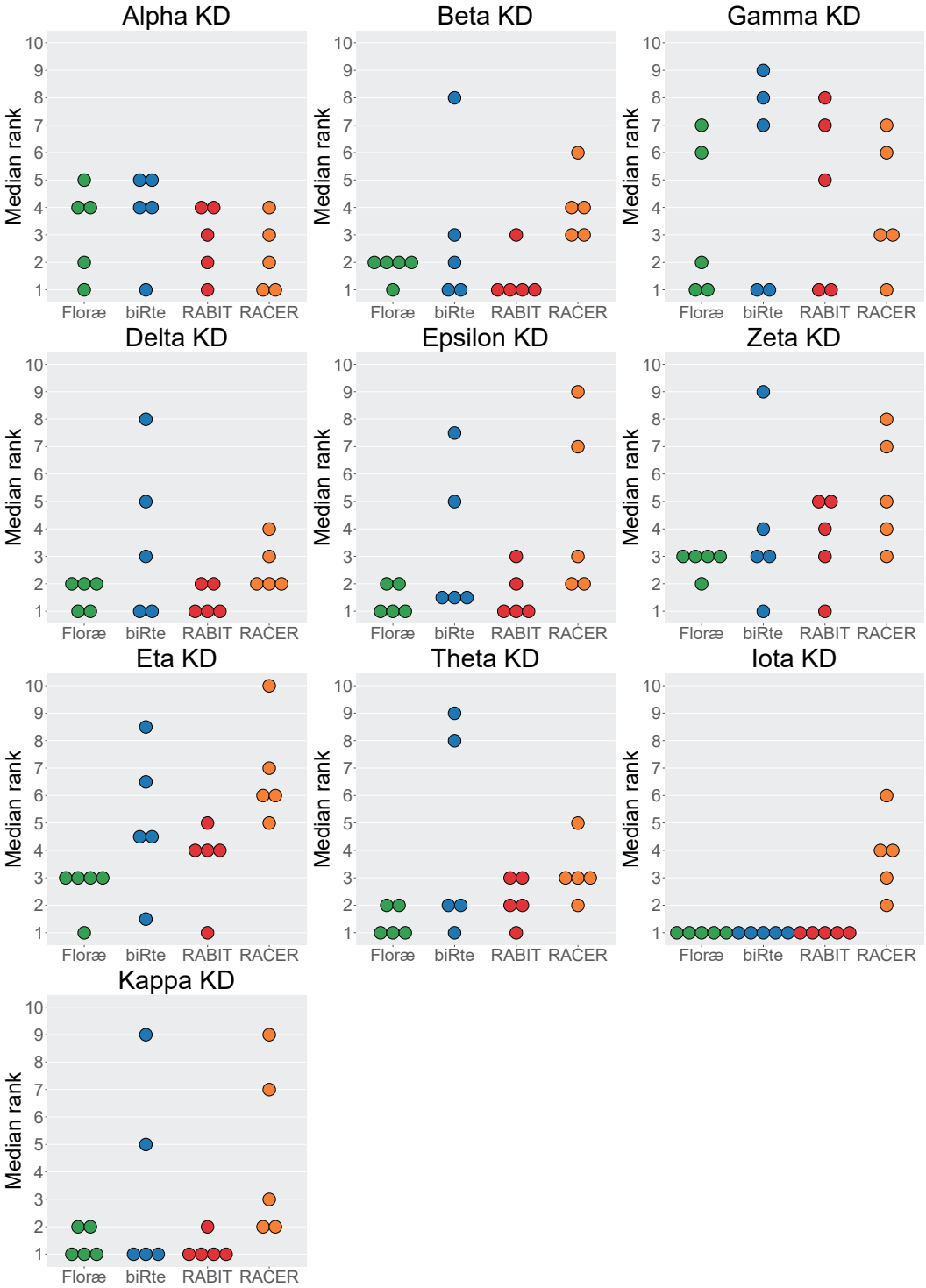
PCA plot for an exemplary simulated WT, KO and KD data set based on network A showing the separation of WT and KO (left panel) respectively WT and KD (right panel) samples. Each WT, KO and KD data set comprises three samples, marked as dots with colors according to the KO/ KD.



A.8. Median ranks for KO and KD TFs per TF

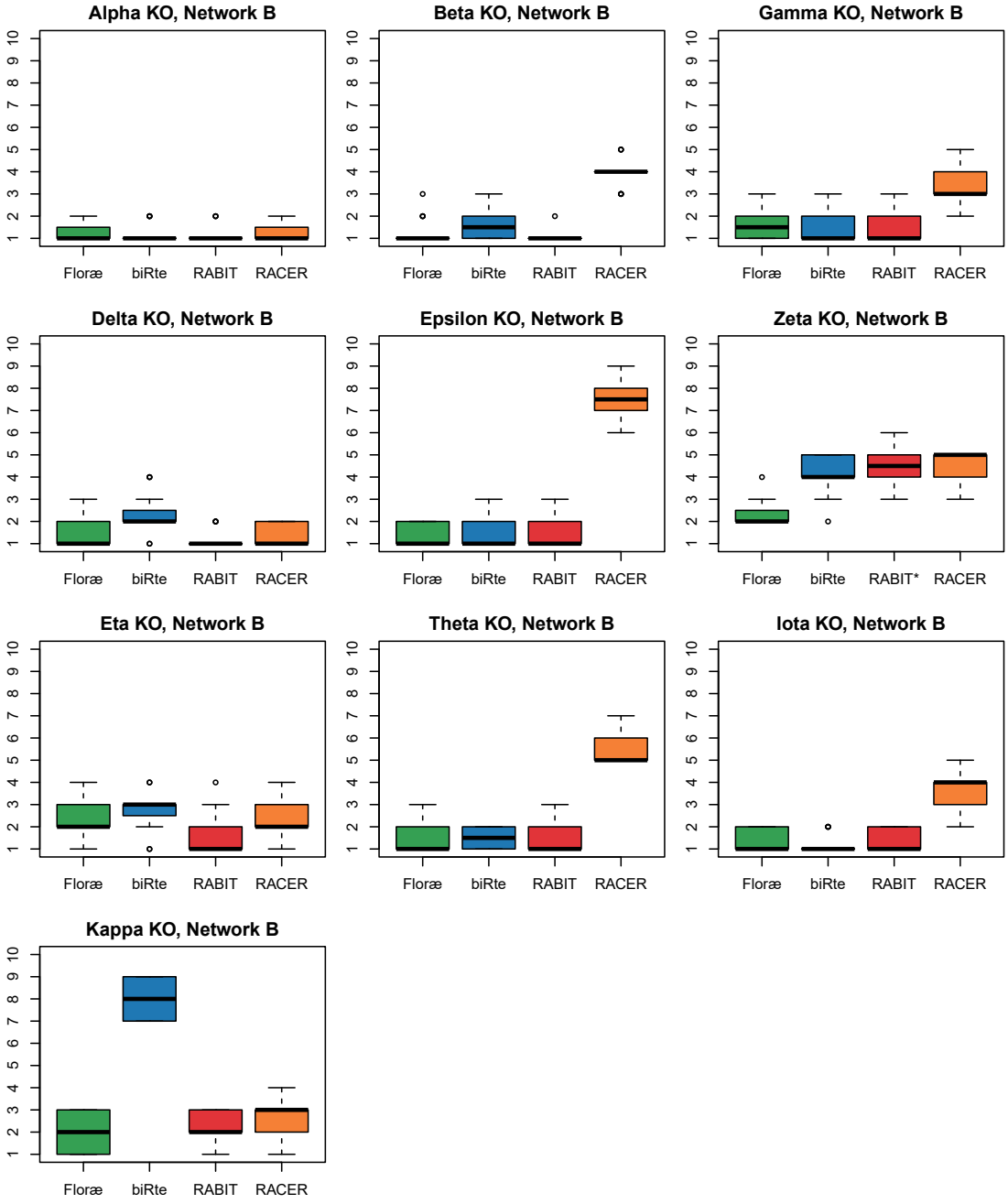
Median ranks of KO and KD TFs over all networks and data sets. Each plot refers to the KO/ KD of a certain TF, the dots represent the median rank that was assigned to the KO/ KD TF over all 20 data sets per method (Floræ: green, biRte: blue, RABIT: red, RACER: orange). Per method, five median ranks are shown (one per network).

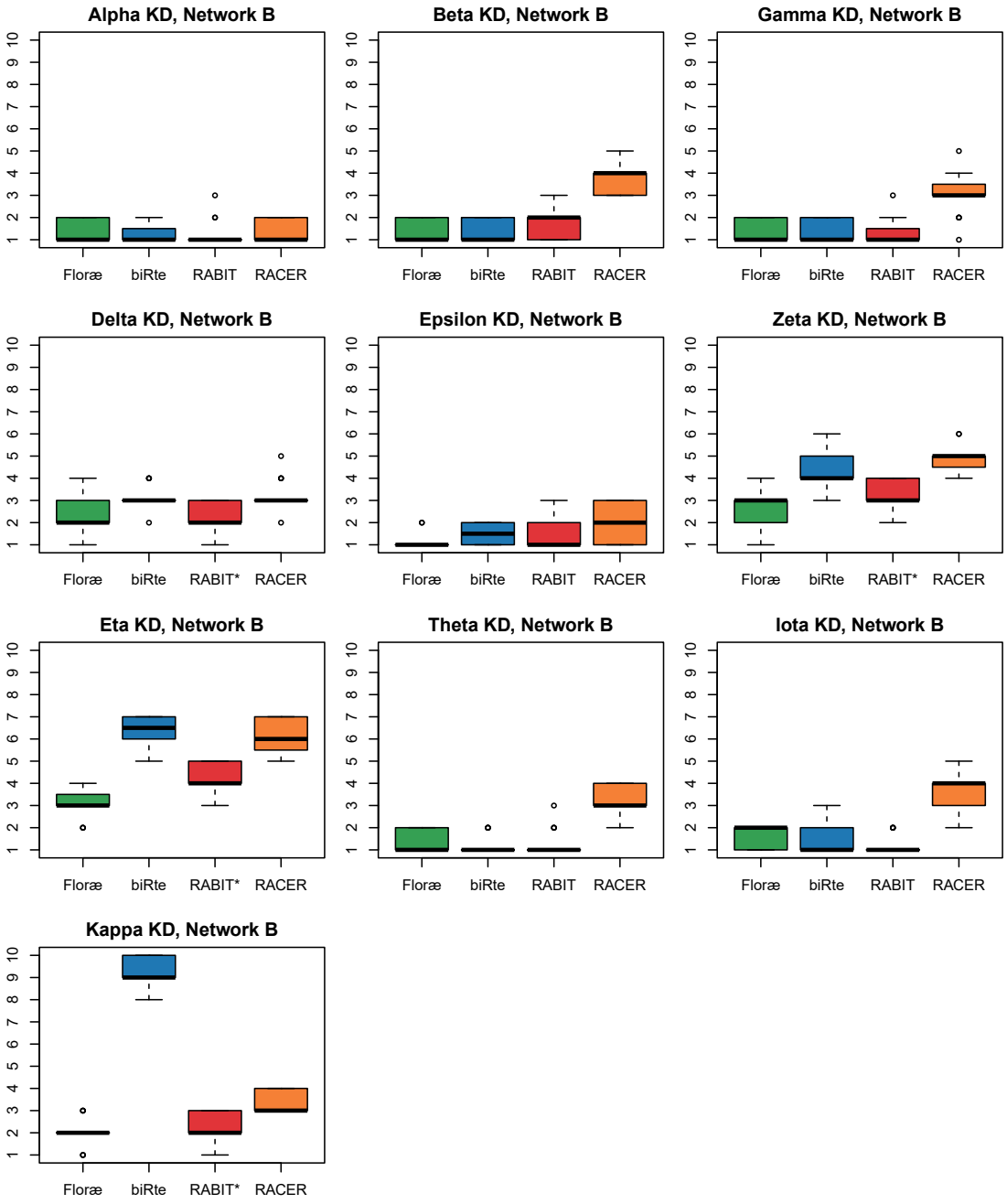

Alpha KO
Beta KO
Gamma KO
Delta KO
Epsilon KO
Zeta KO
Eta KO
Theta KO
Iota KO
Kappa KO
Median rank
1
2
3
4
5
6
7
8
9
10
Floræ
biRte
RABIT
RACER

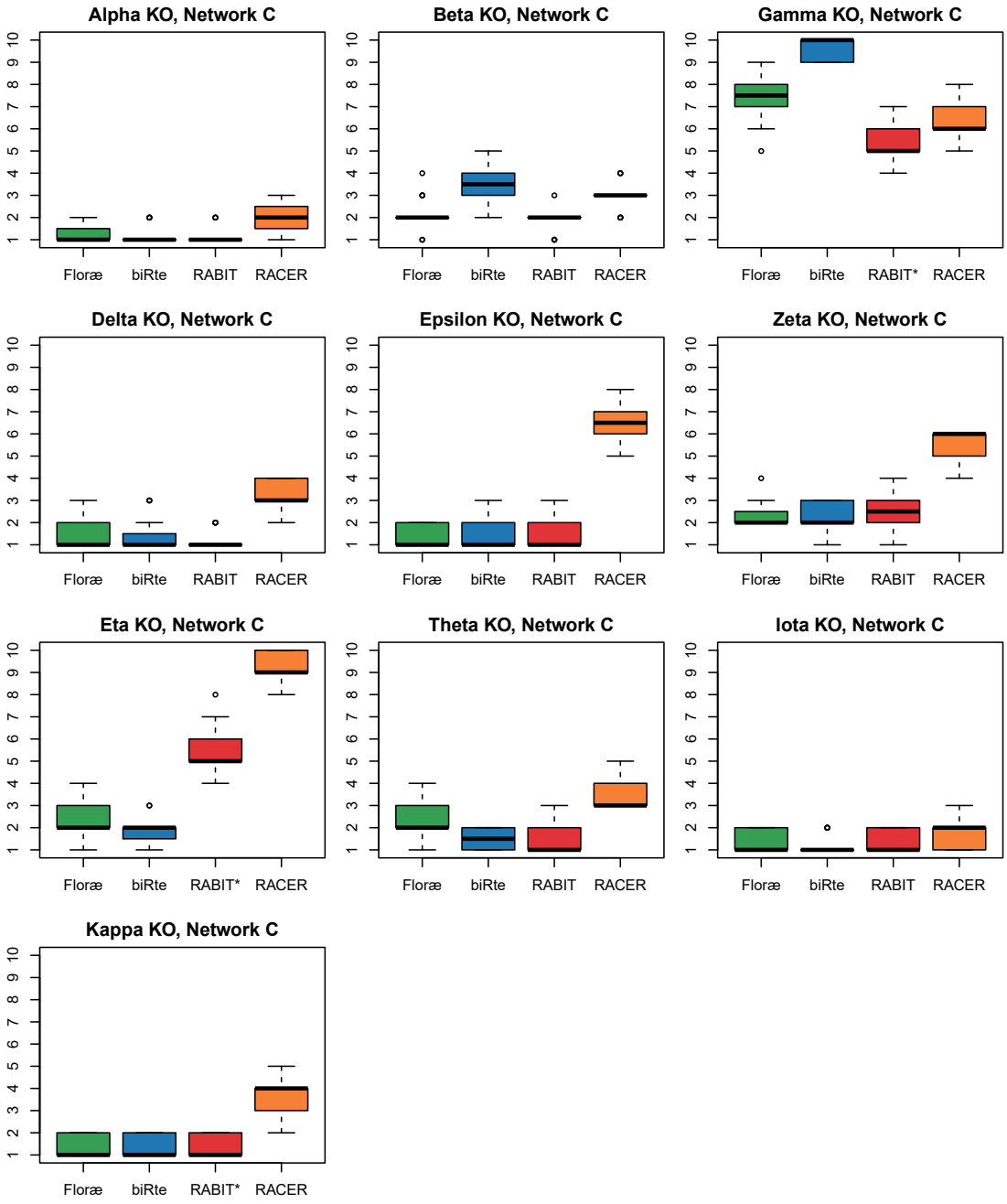

Alpha KD
Beta KD
Gamma KD
Delta KD
Epsilon KD
Zeta KD
Eta KD
Theta KD
Iota KD
Kappa KD
Median rank
1
2
3
4
5
6
7
8
9
10
Floræ
biRte
RABIT
RACER

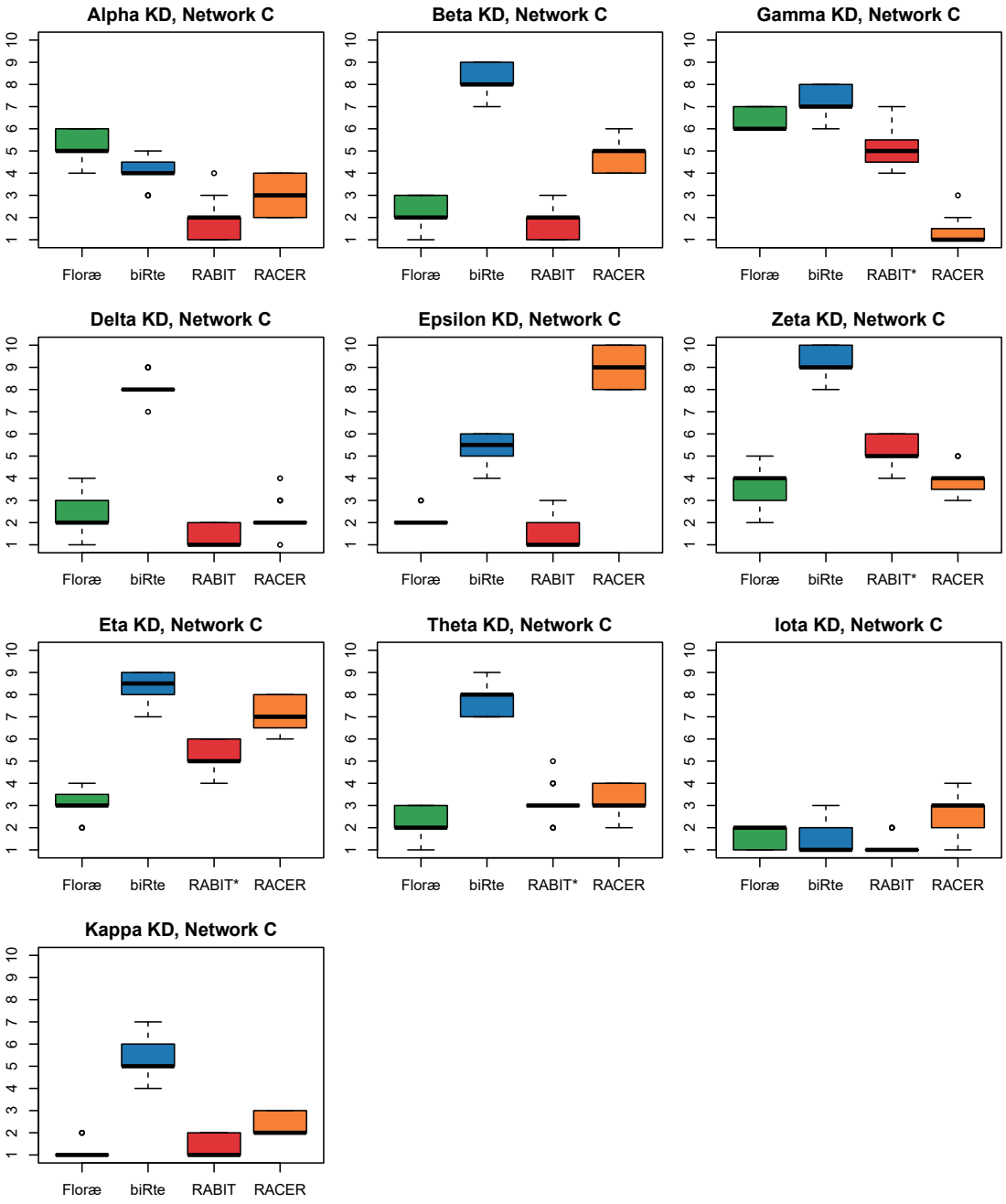
A.9. Ranks of KO and KD TFs for networks B to E

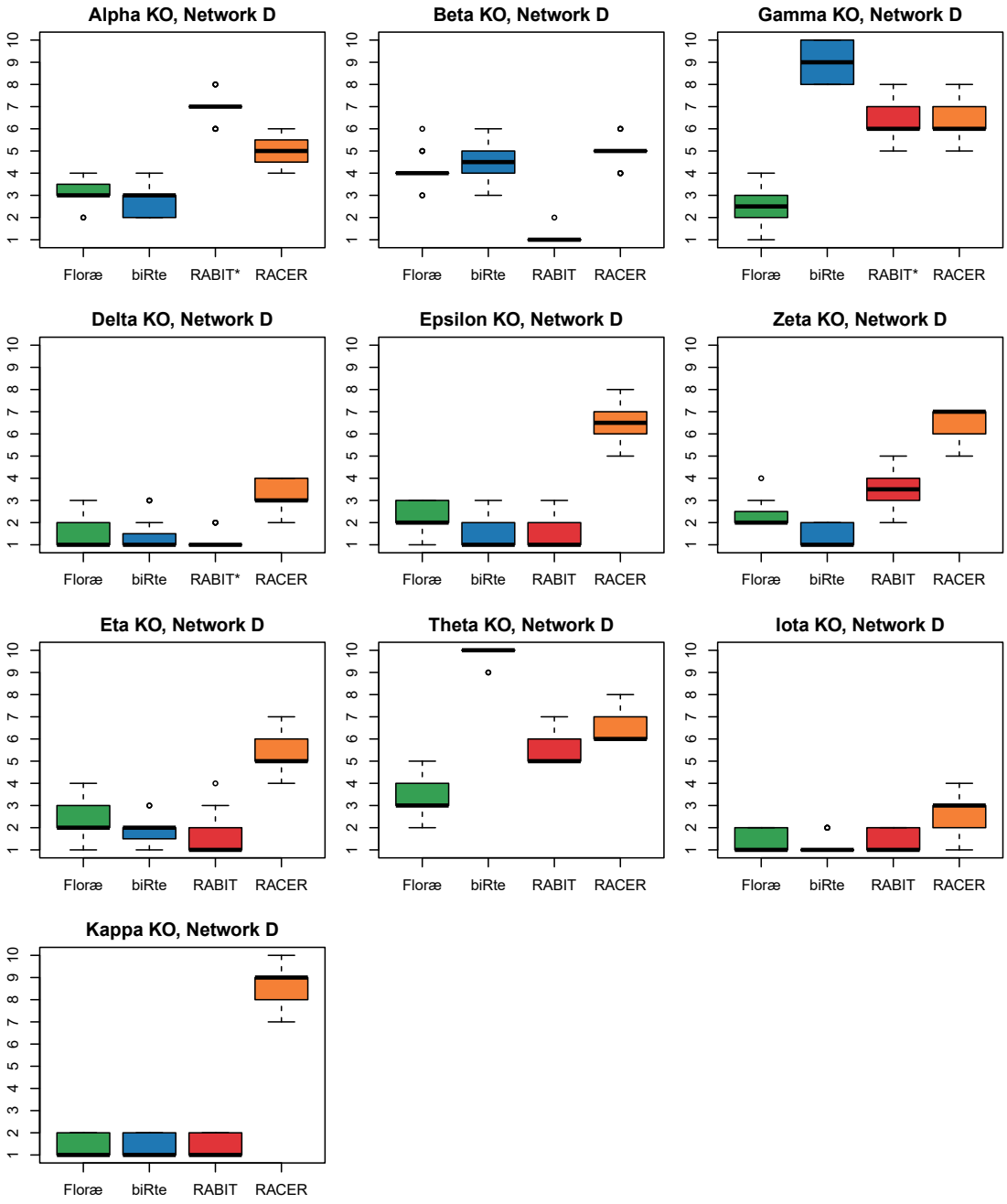
Boxplots showing the TF activity ranks of Floræ (green), biRte (blue), RABIT (red) and RACER (orange) for all ten TF KOs and KDs based on different networks (see headline of each plot), 20 runs of data generation and TF ranking. Median ranks are represented by a bold line, the colored box ranges from 25$^{\text{th}}$ to 75$^{\text{th}}$ percentile, representing the interquartile range.

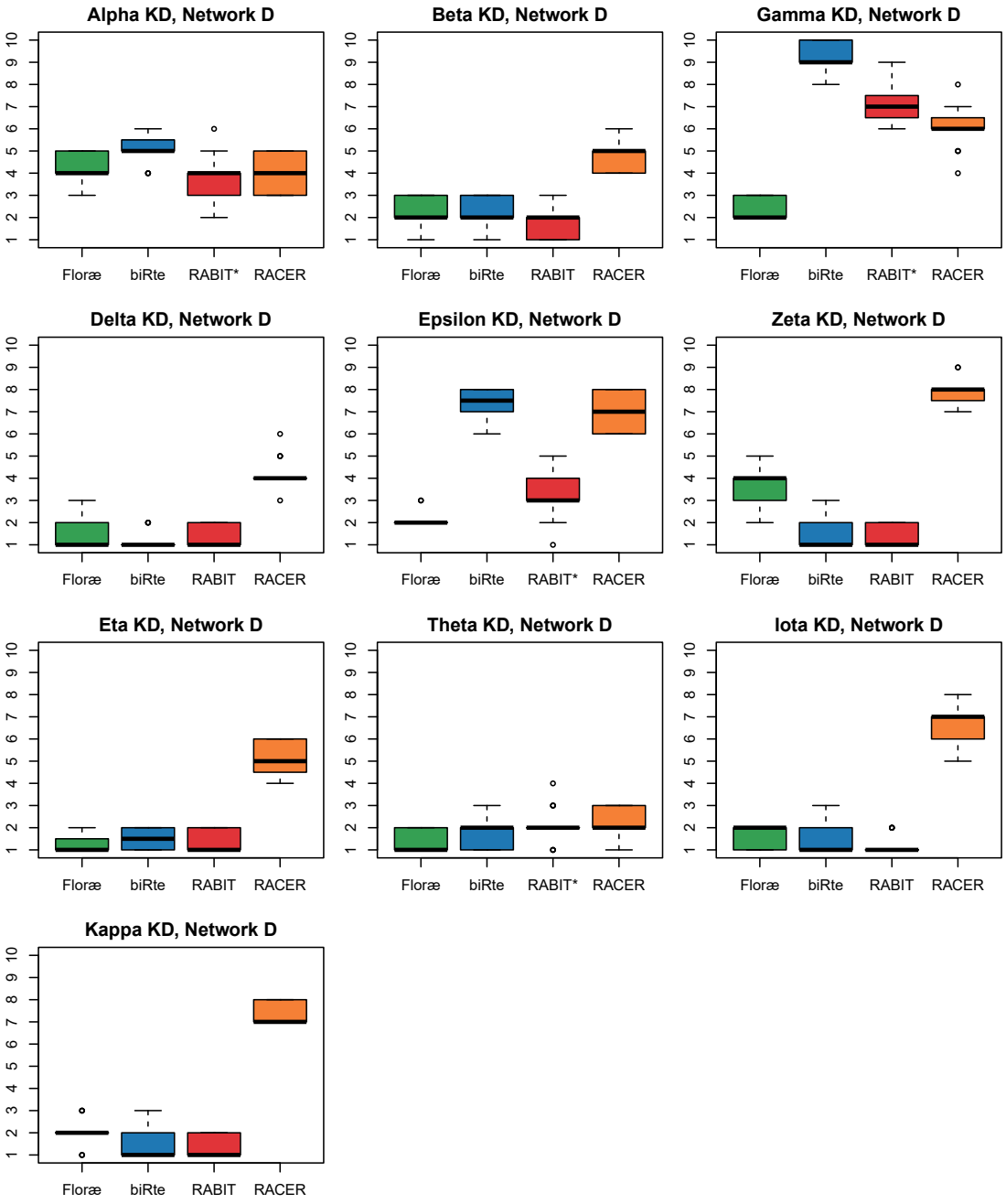


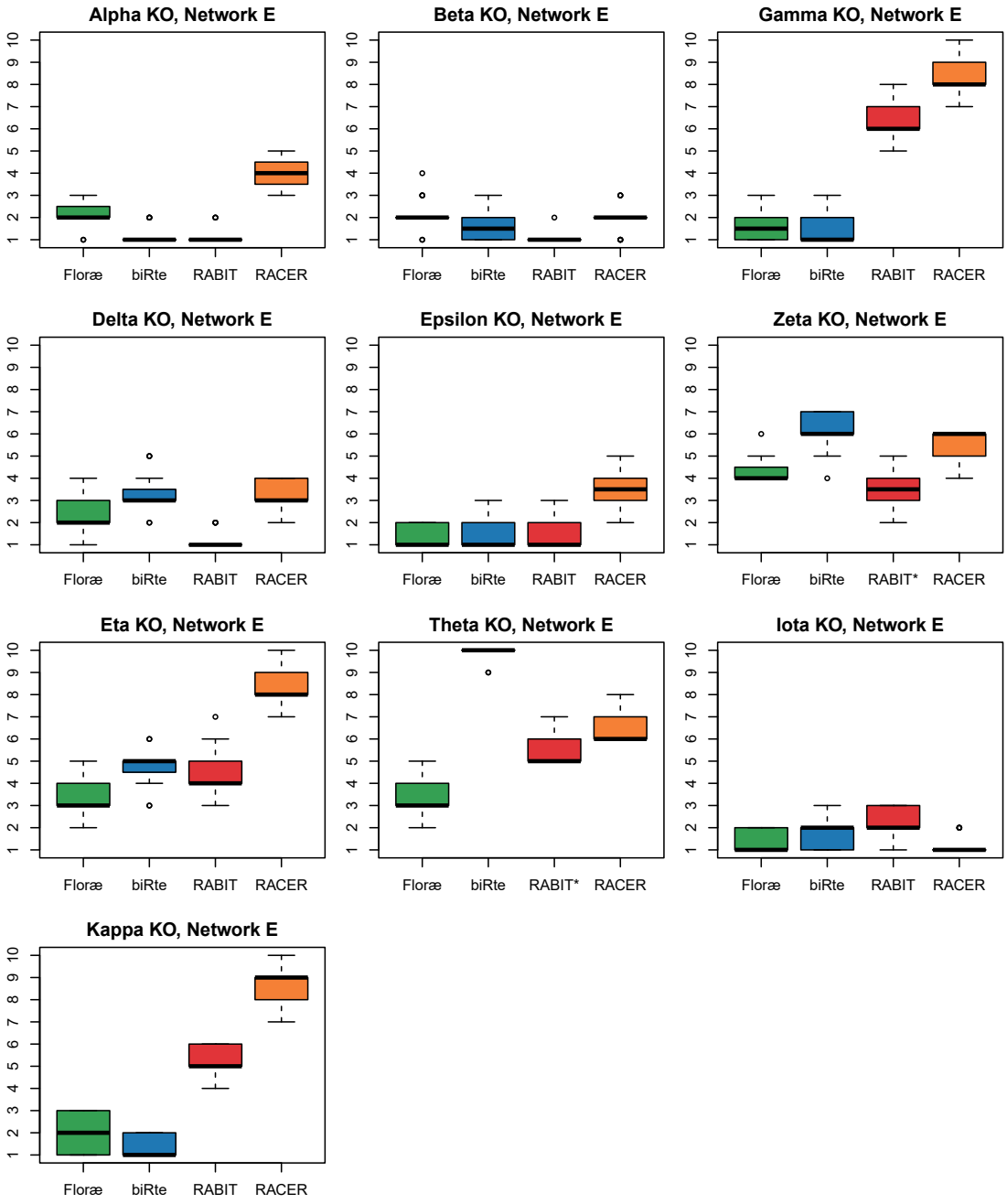

Alpha KD, Network B
Beta KD, Network B
Gamma KD, Network B
Delta KD, Network B
Epsilon KD, Network B
Zeta KD, Network B
Eta KD, Network B
Theta KD, Network B
Iota KD, Network B
Kappa KD, Network B
Floræ
biRte
RABIT
RABIT*
RACER
1 2 3 4 5 6 7 8 9 10

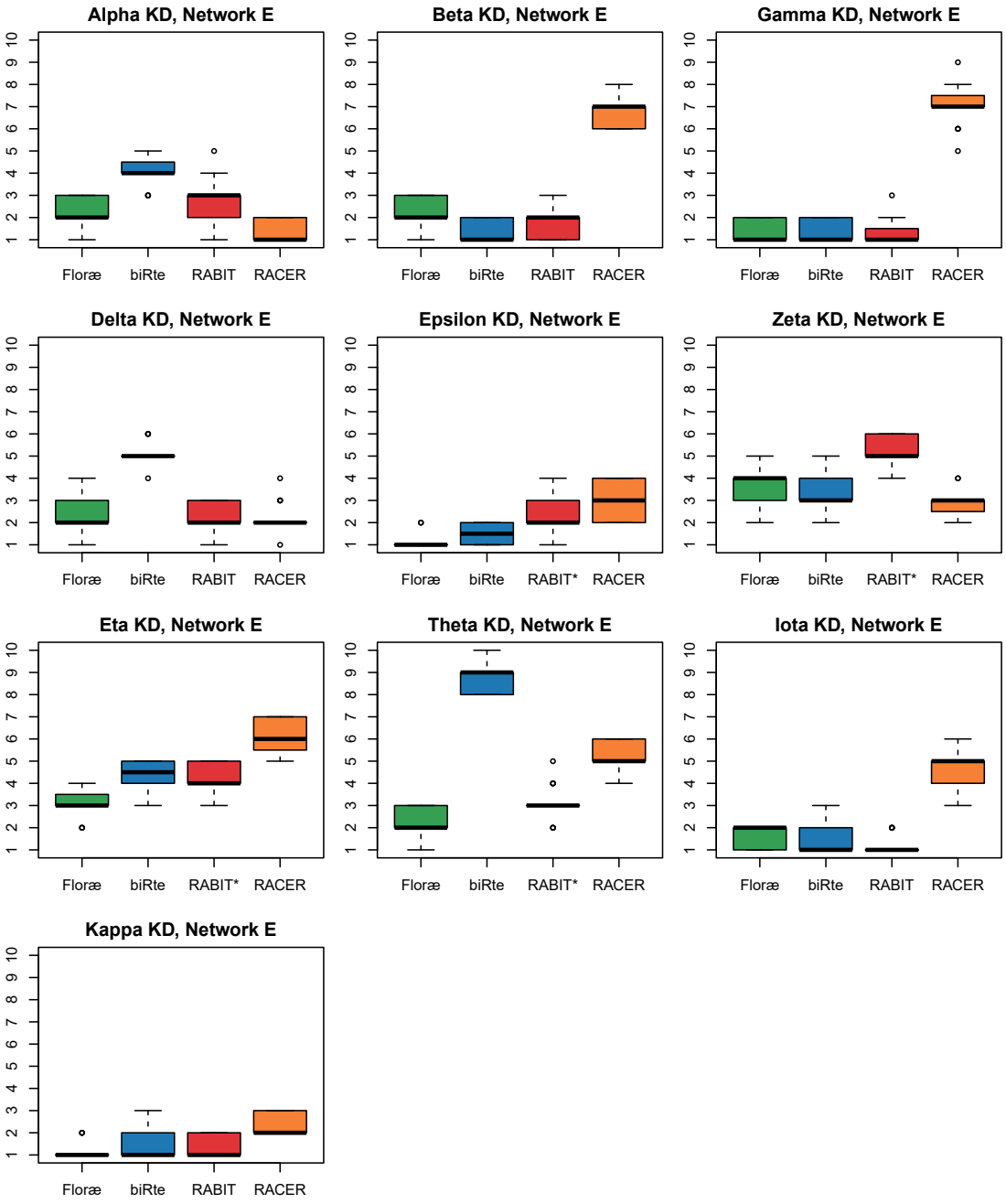

Alpha KO, Network C
Beta KO, Network C
Gamma KO, Network C
Delta KO, Network C
Epsilon KO, Network C
Zeta KO, Network C
Eta KO, Network C
Theta KO, Network C
Iota KO, Network C
Kappa KO, Network C
Floræ
biRte
RABIT
RABIT*
RACER
1 2 3 4 5 6 7 8 9 10


Alpha KD, Network C
Beta KD, Network C
Gamma KD, Network C
Delta KD, Network C
Epsilon KD, Network C
Zeta KD, Network C
Eta KD, Network C
Theta KD, Network C
Iota KD, Network C
Kappa KD, Network C
Floræ
biRte
RABIT
RABIT*
RACER


Alpha KO, Network D
Beta KO, Network D
Gamma KO, Network D
Delta KO, Network D
Epsilon KO, Network D
Zeta KO, Network D
Eta KO, Network D
Theta KO, Network D
Iota KO, Network D
Kappa KO, Network D
Floræ
biRte
RABIT*
RABIT
RACER
1 2 3 4 5 6 7 8 9 10

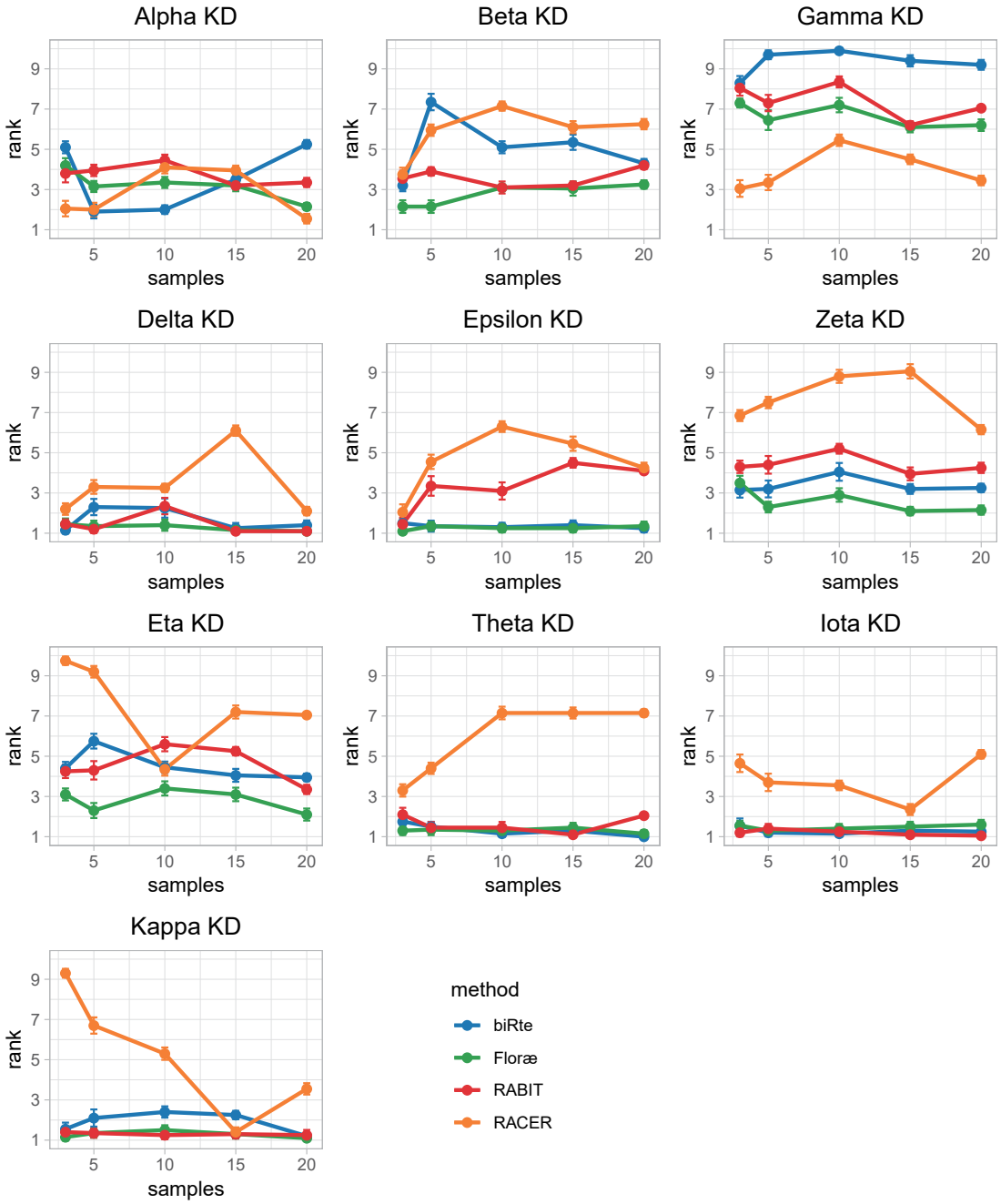

Alpha KD, Network D
Beta KD, Network D
Gamma KD, Network D
Delta KD, Network D
Epsilon KD, Network D
Zeta KD, Network D
Eta KD, Network D
Theta KD, Network D
Iota KD, Network D
Kappa KD, Network D
Floræ
biRte
RABIT*
RABIT
RACER


Alpha KO, Network E
Beta KO, Network E
Gamma KO, Network E
Delta KO, Network E
Epsilon KO, Network E
Zeta KO, Network E
Eta KO, Network E
Theta KO, Network E
Iota KO, Network E
Kappa KO, Network E
Floræ
biRte
RABIT
RABIT*
RACER
1 2 3 4 5 6 7 8 9 10


Alpha KD, Network E
Beta KD, Network E
Gamma KD, Network E
Delta KD, Network E
Epsilon KD, Network E
Zeta KD, Network E
Eta KD, Network E
Theta KD, Network E
Iota KD, Network E
Kappa KD, Network E
Floræ
biRte
RABIT
RABIT*
RACER

A.10. Ranks of KD TFs based on different sample sizes

Scatter Plot indicating mean rank (points) and according standard error of the mean (error bars) of TF activity ranks for all ten knockdown TFs inferred by different methods (biRte: blue, Floræ: green, RABIT: red, RACER: orange) using a varying number of samples (3, 5, 10, 15 and 20) for both wild-type and knockdown data sets. Per sample size, TF activity ranks are calculated on the basis of network A, 20 runs of data generation and TF ranking.



A.11. Randomized networks

Table indicating numbers of cycles in randomized networks (randomization based on network A) and examples for randomized networks with either 10% or 50% changed interactions based on network A. Further, the results for WT vs KD samples (10% randomization), WT vs KO samples (50% randomization) and WT vs KD samples (50% randomization) are given as boxplots.

number of cycles	0	1	2	3	4	5	6
number of cases in 10% randomized networks	0	0	0	0	5	4	1
number of cases in 50% randomized networks	1	2	2	4	1	0	0

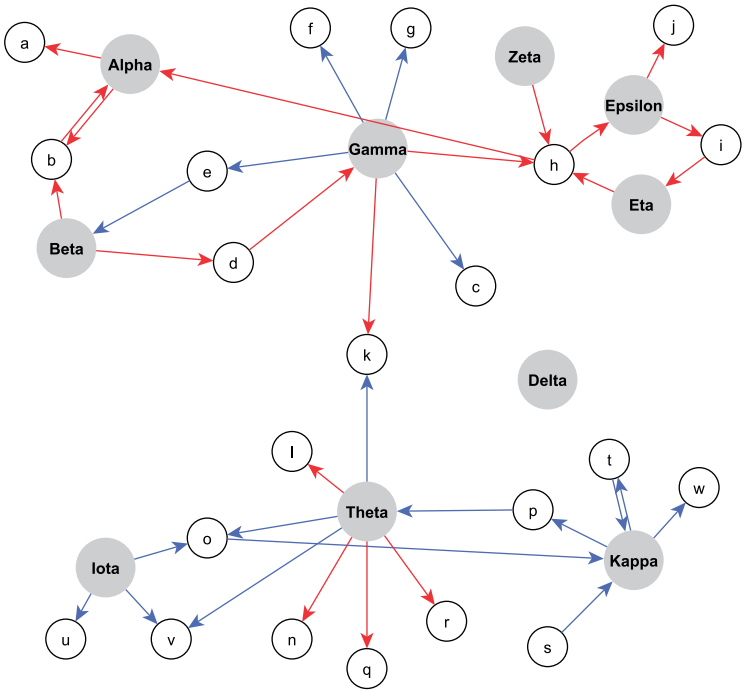


Figure A.1.: 10% randomized edges of network A

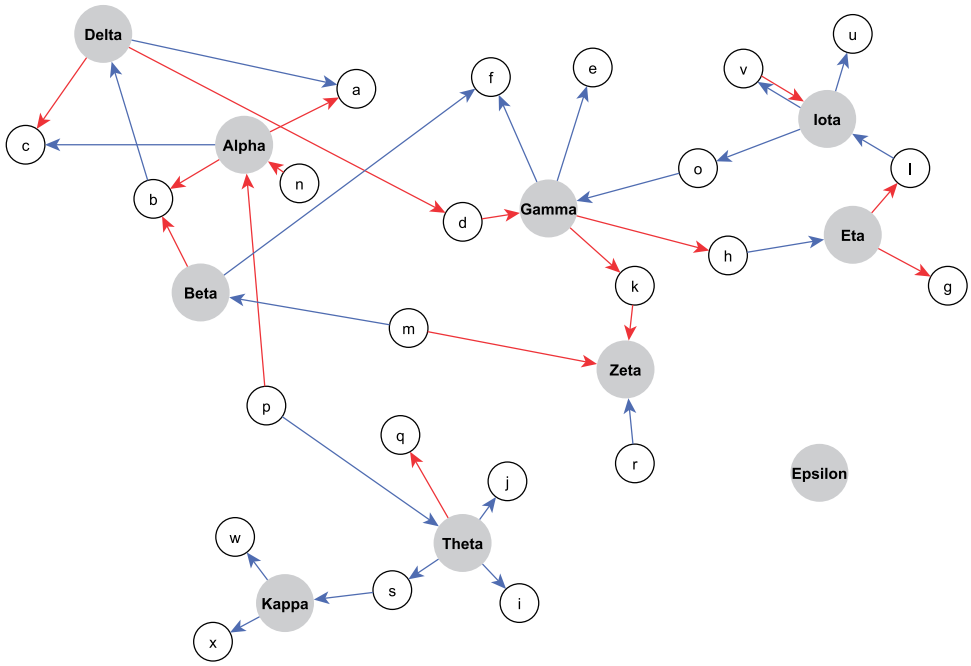


Figure A.2.: 50% randomized edges of network A

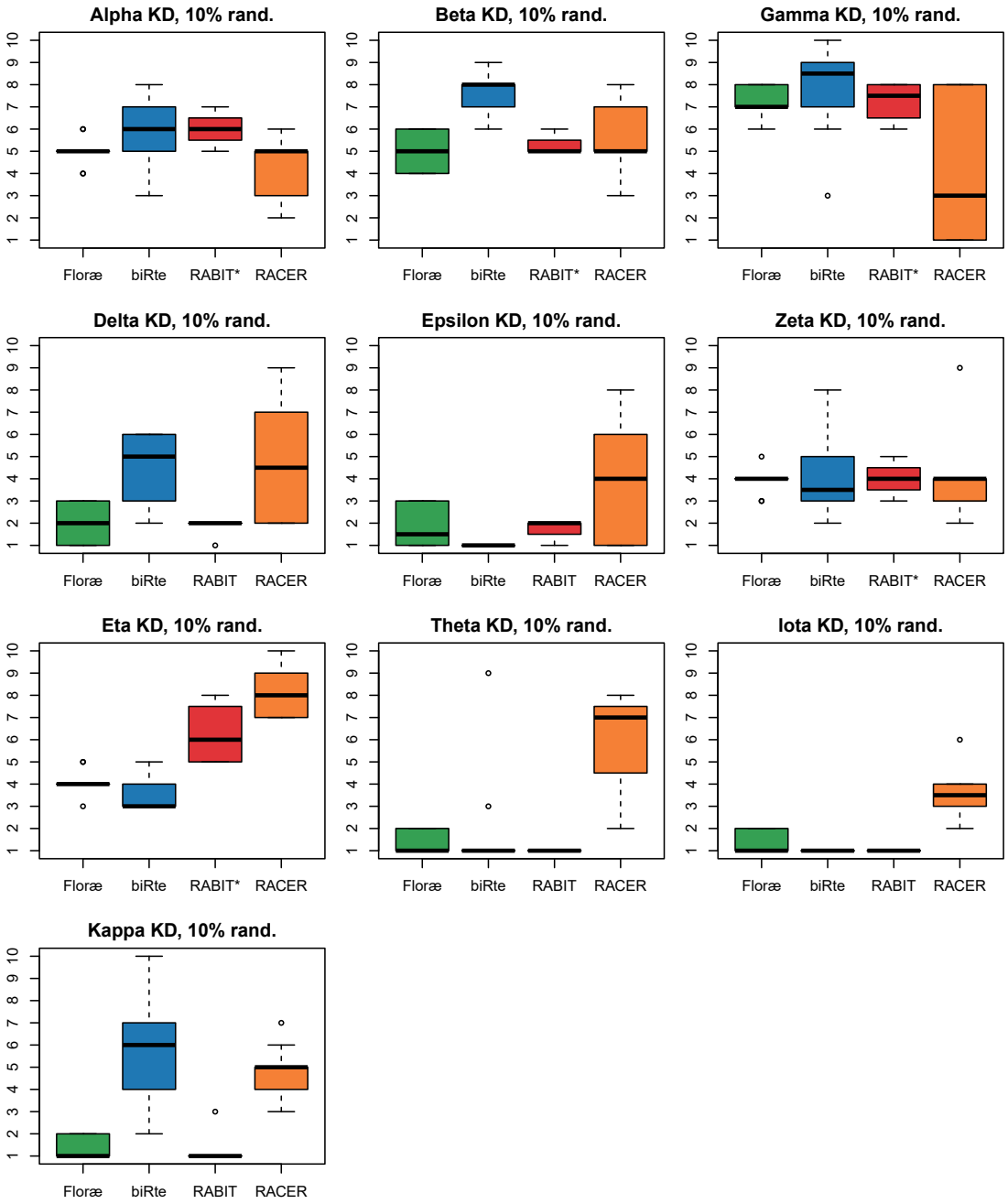


Figure A.3.: Effect of network randomization of network A (10%), WT vs KD samples. Boxplots showing the TF activity ranks of Floræ (green), biRte (blue), RABIT (red) and RACER (orange) for all ten TF knockouts.

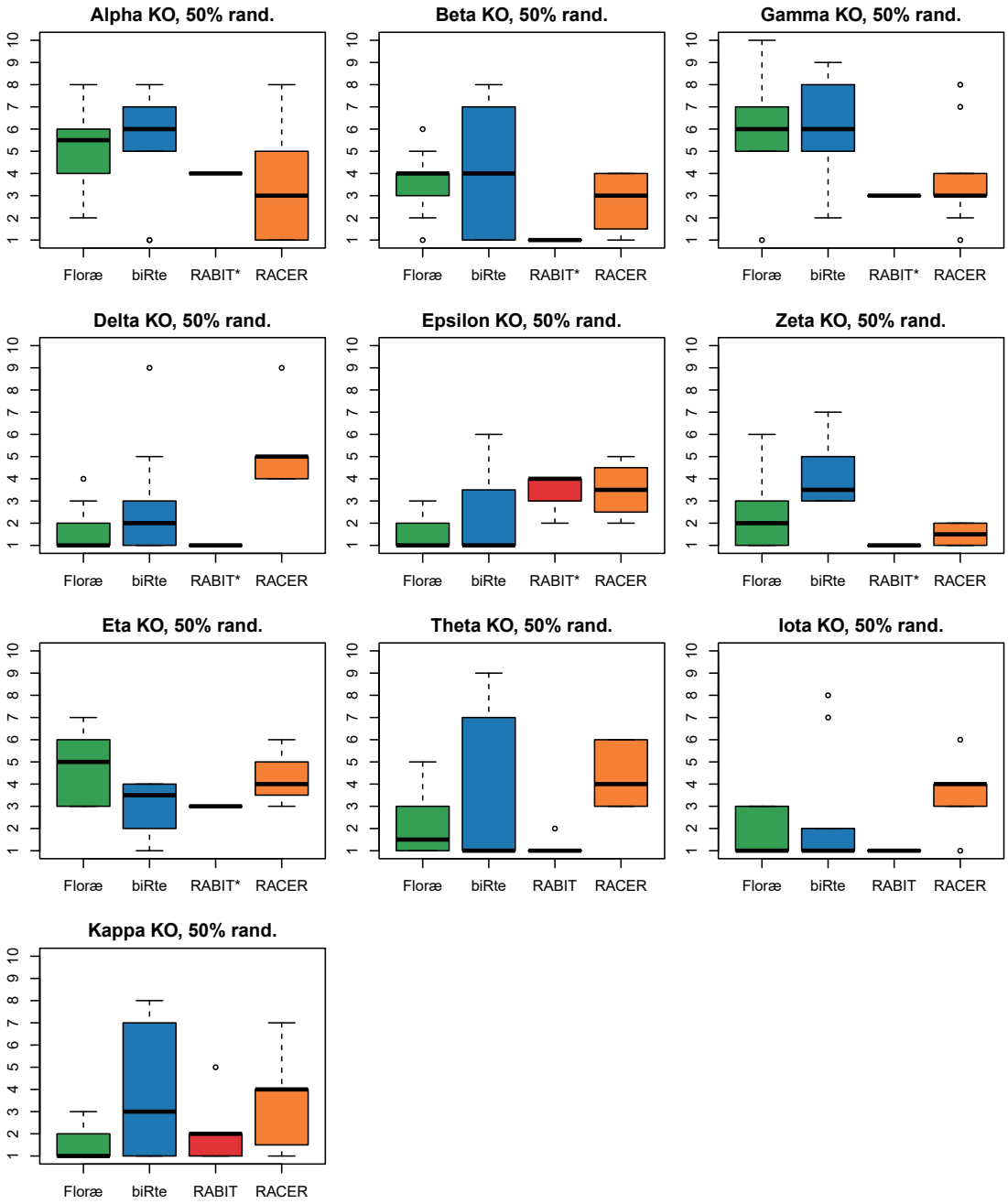


Figure A.4.: Effect of network randomization of network A (50%), WT vs KO samples. Boxplots showing the TF activity ranks of Floræ (green), biRte (blue), RABIT (red) and RACER (orange) for all ten TF knockouts.

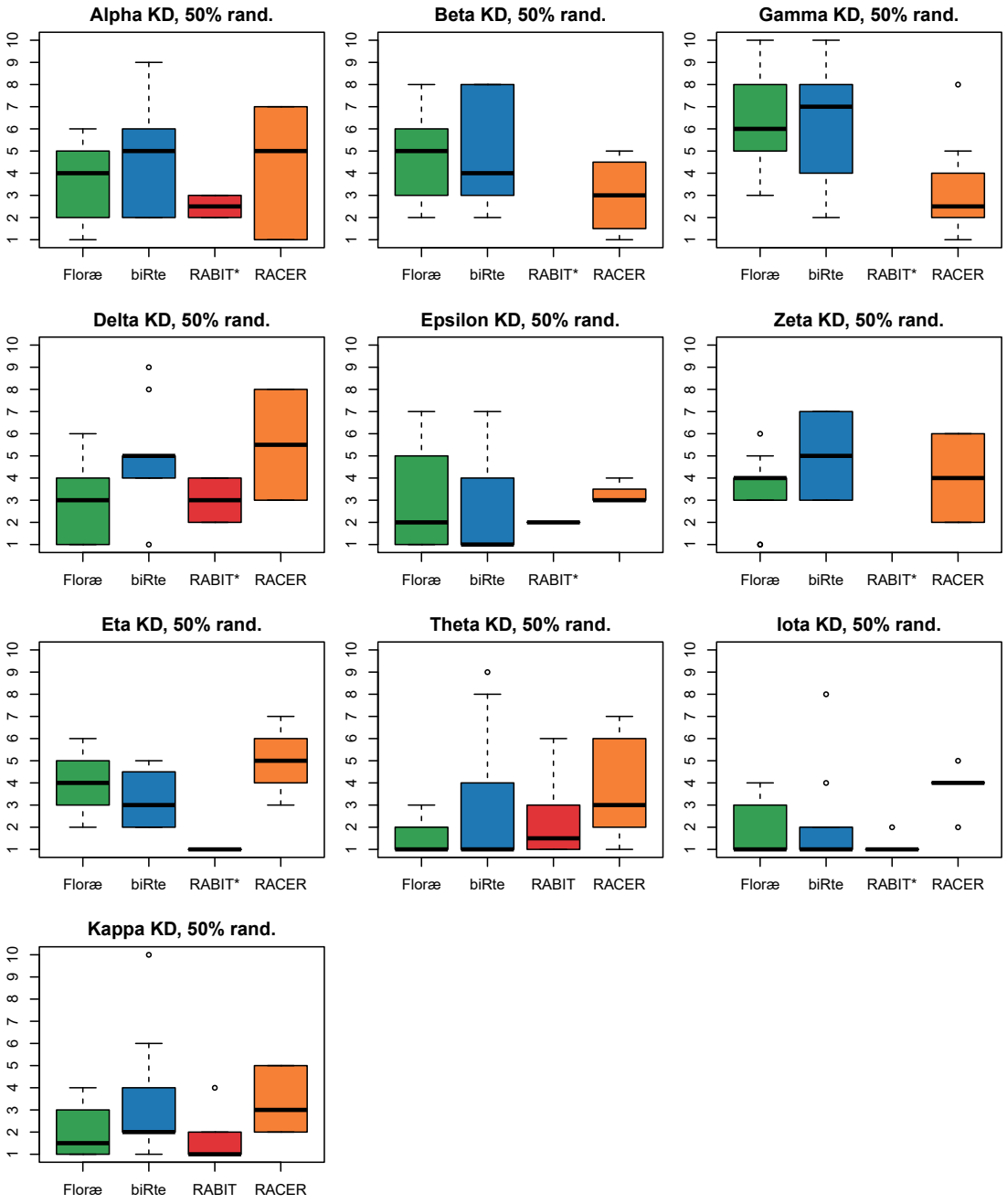


Figure A.5.: Effect of network randomization of network A (50%), WT vs KD samples. Boxplots showing the TF activity ranks of Floræ (green), biRte (blue), RABIT (red) and RACER (orange) for all ten TF knockouts.

Bibliography

K. L. Abbott, E. T. Nyre, J. Abrahante, Y. Y. Ho, R. I. Vogel, and T. K. Starr. The candidate cancer gene database: A database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Research*, 43:D844–D848, 2015.

N. Aceto, N. Sausgruber, H. Brinkhaus, D. Gaidatzis, G. Martiny-Baron, G. Mazzarol, S. Confalonieri, M. Quarto, G. Hu, P. J. Balwierz, M. Pachkov, S. J. Elledge, E. Van Nimwegen, M. B. Stadler, and M. Bentires-Alj. Tyrosine phosphatase SHP2 promotes breast cancer progression and maintains tumor-initiating cells via activation of key transcription factors and a positive feedback signaling loop. *Nature Medicine*, 18(4):529–537, 2012.

R. Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(21): 4947–4957, 2005.

B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. C. Raff, K. Roberts, and P. Walter. *Essential cell biology*. Garland Science, New York, 2014.

D. J. Allocco, I. S. Kohane, and A. J. Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC bioinformatics*, 25:5–18, 2004.

U. Alon. Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 8:450–461, 2007.

M. J. Alvarez, P. Sumazin, P. Rajbhandari, and A. Califano. Correlating measurements across samples improves accuracy of large-scale expression profile experiments. *Genome Biology*, 10, 2009.

M. J. Alvarez, Y. Shen, F. M. Giorgi, A. Lachmann, B. B. Ding, B. Hilda Ye, and A. Califano. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics*, 48(8):838–847, 2016.

A. Aribi, G. Borthakur, F. Ravandi, J. Shan, J. Davisson, J. Cortes, and H. Kantarjian. Activity of decitabine, a hypomethylating agent, in chronic myelomonocytic leukemia. *Cancer*, 109(4):713–717, 2007.

F. Atger, C. Gobet, J. Marquis, E. Martin, J. Wang, B. Weger, G. Lefebvre, P. Descombes, F. Naef, and F. Gachon. Circadian and feeding rhythms differentially affect rhythmic mRNA transcription and translation in mouse liver. *Proceedings of the National Academy of Sciences*, 112:E6579–E6588, 2015.

A. C. Babtie, P. Kirk, and M. P. Stumpf. Topological sensitivity analysis for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, 111(52):18507–18512, 2014.

M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283–291, 2004.

S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1): 77–120, 2017.

P. Balwierz, M. Pachkov, P. Arnold, A. J. Gruber, Z. Mihaela, and E. van Nimwegen. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome research*, 24(5):869–884, 2014.

M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Molecular systems biology*, 3:78, 2007.

M. S. Bazaraa, J. J. Jarvis, and H. D. Sherali. *Linear programming and network flows: Fourth edition*. John Wiley & Sons, 2011.

D. R. Beers, W. Zhao, J. Wang, X. Zhang, S. Wen, D. Neal, J. R. Thonhoff, A. S. Alsuliman, E. J. Shpall, K. Rezvani, and S. H. Appel. ALS patients' regulatory T lymphocytes are dysfunctional, and correlate with disease progression rate and severity. *JCI Insight*, 2(5), 2017.

E. Berchtold, G. Csaba, and R. Zimmer. Evaluating transcription factor activity changes by scoring unexplained target genes in expression data. *PLoS ONE*, 11(10): 1–16, 2016.

N. Berestovsky and L. Nakhleh. An Evaluation of Methods for Inferring Boolean Networks from Time-Series Data. *PLoS ONE*, 8(6), 2013.

M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, and C. S. Haley. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports*, 5:10312, 2015.

M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, and L. Milanesi. Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinformatics*, 17(2), 2016.

A. S. Bhagwat and C. R. Vakoc. Targeting Transcription Factors in Cancer. *Trends in Cancer*, 1(1):53–65, 2015.

X. Bisteau, M. J. Caldez, and P. Kaldis. The complex relationship between liver cancer and the cell cycle: A story of multiple regulations. *Cancers*, 6(1):79–111, 2014.

B. A. Boghigian, H. Shi, K. Lee, and B. A. Pfeifer. Utilizing elementary mode analysis, pathway thermodynamics, and a genetic algorithm for metabolic flux determination and optimal metabolic network design. *BMC Systems Biology*, 4(1), 2010.

H. Bolouri. Modeling genomic regulatory networks with big data. *Trends in Genetics*, 30(5):182–191, 2014.

H. Bolouri and E. H. Davidson. Modeling transcriptional regulatory networks. *BioEssays*, 24(12):1118–1129, 2002.

M. J. Bonder, R. Luijk, D. V. Zhernakova, M. Moed, P. Deelen, M. Vermaat, M. Van Iterson, F. Van Dijk, M. Van Galen, J. Bot, R. C. Slieker, P. M. Jhamai, M. Verbiest, H. E. D. Suchiman, M. Verkerk, R. Van Der Breggen, J. Van Rooij, N. Lakenberg, W. Arindrarto, S. M. Kielbasa, I. Jonkers, P. Van't Hof, I. Nooren, M. Beekman, J. Deelen, D. Van Heemst, A. Zhernakova, E. F. Tigchelaar, M. A. Swertz, A. Hofman, A. G. Uitterlinden, R. Pool, J. Van Dongen, J. J. Hottenga, C. D. Stehouwer, C. J. Van Der Kallen, C. G. Schalkwijk, L. H. Van Den Berg, E. W. Van Zwet, H. Mei, Y. Li, M. Lemire, T. J. Hudson, P. E. Slagboom, C. Wijmenga, J. H. Veldink, M. M. Van Greevenbroek, C. M. Van Duijn, D. I. Boomsma, A. Isaacs, R. Jansen, J. B. Van Meurs, P. A. Hoen't, L. Franke, and B. T. Heijmans. Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*, 49(1):131–138, 2017.

A. Bonnaffoux, U. Herbach, A. Richard, A. Guillemin, S. Gonin-Giraud, P. A. Gros, and O. Gandrillon. WASABI: A dynamic iterative framework for gene regulatory network inference. *BMC Bioinformatics*, 20(1), 2019.

A. Boorsma, B. C. Foat, D. Vis, F. Klis, and H. J. Bussemaker. T-profiler: Scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Research*, 33(SUPPL. 2), 2005.

A. L. Boulesteix and K. Strimmer. Predicting transcription factor activities from combined analysis of microarray and ChIP data: A partial least squares approach. *Theoretical Biology and Medical Modelling*, 2005.

O. Brandman and T. Meyer. Feedback loops shape cellular signals in space and time. *Science*, 322(5900):390–395, 2008.

J. Brandt, M. Bux, and U. Leser. Cuneiform: A Functional Language for Large Scale Scientific Data Analysis. *Proceedings of the Workshops of the EDBT/ICDT*, 1330: 17–26, 2015.

A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME) - Toward standards for microarray data. *Nature Genetics*, 29(4):365–371, 2001.

M. R. Brent. Past Roadblocks and New Opportunities in Transcription Factor Network Mapping. *Trends in Genetics*, 32(11):736–750, 2016.

D. M. Budden and E. J. Crampin. Information theoretic approaches for inference of biological networks from continuous-valued data. *BMC Systems Biology*, 10(1):89, 2016.

A. J. Butte and I. S. Kohane. Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 418–429, 2013.

M. Bux, J. Brandt, C. Lipka, K. Hakimazadeh, J. Dowling, and U. Leser. SAASFEE: Scalable Scientific Workflow Execution Engine. *Proceedings of the VLDB*, 8(12): 1892–1895, 2015.

M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape, A. Califano, and A. Iavarone. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463:318–325, 2010.

L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, and Z. Zakaria. A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, 48(1):55–65, 2014.

A. Chatterjee, P. A. Stockwell, E. J. Rodger, and I. M. Morison. Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Research*, 40(10):e79, 2012.

T. Chekouo, F. C. Stingo, J. D. Doecke, and K. A. Do. miRNA-target gene regulatory networks: A Bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics*, 71(2):428–438, 2015.

H. Chen, H. Liu, and G. Qing. Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduction and Targeted Therapy*, 3(1), 2018.

Y. Chen, M. Widschwendter, and A. E. Teschendorff. Systems-epigenomics inference of transcription factor activity implicates aryl-hydrocarbon-receptor inactivation as a key event in lung cancer development. *Genome Biology*, 18(1), 2017.

V. G. Cheung, M. Morley, F. Aguilar, A. Massimi, R. Kucherlapati, and G. Childs. Making and reading microarrays. *Nature Genetics*, 21(1S):19, 1999.

E. Chipumuro, E. Marco, C. L. Christensen, N. Kwiatkowski, T. Zhang, C. M. Hatheway, B. J. Abraham, B. Sharma, C. Yeung, A. Altabef, A. Perez-Atayde, K. K. Wong, G. C. Yuan, N. S. Gray, R. A. Young, and R. E. George. CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. *Cell*, 159(5):1126–1139, 2014.

S. Chrétien and A. O. Hero. On EM algorithms and their proximal generalizations. *ESAIM - Probability and Statistics*, 12:308–326, 2008.

J. H. Chu, S. T. Weiss, V. J. Carey, and B. A. Raby. A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC Systems Biology*, 3(1), 2009.

C. Cillo, A. Faiella, M. Cantile, and E. Boncinelli. Homeobox genes and cancer. *Experimental Cell Research*, 248(1):1–9, 1999.

C. R. Clapier and B. R. Cairns. The Biology of Chromatin Remodeling Complexes. *Annual Review of Biochemistry*, 78(1):273–304, 2009.

D. J. Clarke, M. V. Kuleshov, B. M. Schilder, D. Torre, M. E. Duffy, A. B. Keenan, A. Lachmann, A. S. Feldmann, G. W. Gundersen, M. C. Silverstein, Z. Wang, and A. Ma'Ayan. EXpression2Kinases (X2K) Web: Linking expression signatures to upstream cell signaling networks. *Nucleic Acids Research*, 46(W1):W171–W179, 2018.

C. V. Clevenger. Roles and regulation of stat family transcription factors in human breast cancer. *American Journal of Pathology*, 165(5):1449–1460, 2004.

F. S. Collins, M. Morgan, and A. Patrinos. The Human Genome Project: Lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.

I. Compan and D. Touati. Anaerobic activation of arcA transcription in Escherichia coli: roles of Fnr and ArcA. *Molecular Microbiology*, 11:955–964, 1994.

L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. An improved algorithm for matching large graphs. In *3rd IAPR-TC15 workshop on graph-based representations in pattern recognition*, pages 149–159, 2001.

F. J. Couch, M. R. Johnson, K. G. Rabe, K. Brune, M. de Andrade, M. Goggins, H. Rothenmund, S. Gallinger, A. Klein, G. M. Petersen, and R. H. Hruban. The prevalence of BRCA2 mutations in familial pancreatic cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 16(2): 342–346, 2007.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429:92–96, 2004.

C. V. Dang. MYC on the path to cancer. *Cell*, 149(1):22–35, 2012.

C. O. Daub, R. Steuer, J. Selbig, and S. Kloska. Estimating mutual information using B-spline functions - An improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5:118, 2004.

P. K. Davidsen, N. Turan, S. Egginton, and F. Falciani. Multi-level functional genomics data integration as a tool for understanding physiology: A network perspective. *Journal of applied physiology*, 120(3):297–309, 2016.

H. De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

E. de Wit and W. de Laat. A decade of 3C technologies: Insights into nuclear organization. *Genes and Development*, 26(1):11–24, 2012.

F. M. Delgado and F. Gómez-Vela. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artificial Intelligence in Medicine*, 95:133–145, 2019.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm . *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Y. Deng, H. Zenil, J. Tegnér, and N. A. Kiani. HiDi: An efficient reverse engineering schema for large-scale dynamic regulatory network reconstruction using adaptive differentiation. *Bioinformatics*, 33(24):3964–3972, 2017.

P. D'Haeseleer, S. Liang, and R. Somogyi. Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

B. Di Camillo, G. Toffolo, and C. Cobelli. A gene network simulator to assess reverse engineering algorithms. *Annals of the New York Academy of Sciences*, 1158:125–142, 2009.

E. Di Zanni, G. Bianchi, R. Ravazzolo, L. Raffaghello, I. Ceccherini, and T. Bachetti. Targeting of PHOX2B expression allows the identification of drugs effective in counteracting neuroblastoma cell growth. *Oncotarget*, 8(42):72133–72146, 2017.

C. Dupont, D. R. Armant, and C. A. Brenner. Epigenetics: Definition, mechanisms and clinical perspective. *Seminars in Reproductive Medicine*, 27(5):351–357, 2009.

S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, 4(8):1184–91, 2009.

R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1): 207–10, 2002.

O. Elemento and S. Tavazoie. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biology*, 6:R18, 2005.

D. C. Ellwanger, J. F. Leonhardt, and H. W. Mewes. Large-scale modeling of condition-specific gene regulatory networks by information integration and inference. *Nucleic acids research*, 42(21), 2014.

J. Ernst, O. Vainas, C. T. Harbison, I. Simon, and Z. Bar-Joseph. Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, 3, 2007.

J. Ernst, H. L. Plasterer, I. Simon, and Z. Bar-Joseph. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Research*, 20(4):526–536, 2010.

A. Esquela-Kerscher and F. J. Slack. Oncomirs - microRNAs with a role in cancer. *Nature reviews. Cancer*, 6(4):259–69, 2006.

M. Esteller. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, 8(4):286–298, 2007.

J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):0054–0066, 2007.

D. Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904, 2012.

X. Fang, A. Sastry, N. Mih, D. Kim, J. Tan, J. T. Yurkovich, C. J. Lloyd, Y. Gao, L. Yang, and B. O. Palsson. Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. *Proceedings of the National Academy of Sciences*, 114:10286–10291, 2017.

T. A. Farazi, J. I. Spitzer, P. Morozov, and T. Tuschl. mirnas in human cancer. *The Journal of Pathology*, 223(2):102–115, 2011.

D. Fazekas, M. Koltai, D. Türei, D. Módos, M. Pálfy, Z. Dúl, L. Zsákai, M. Szalay-Beko, K. Lenti, I. J. Farkas, T. Vellai, P. Csermely, and T. Korcsmáros. SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC systems biology*, 7, 2013.

J. Fluck and M. Hofmann-Apitius. Text mining for systems biology. *Drug Discovery Today*, 19(2):140–144, 2014.

A. Ford. Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems. *International Journal of Sustainability in Higher Education*, 1(1):56–57, 2000.

D. Fouskakis and D. Draper. Stochastic optimization: A review. *International Statistical Review*, 70(3):315–349, 2002.

M. F. Fraga, E. Ballestar, A. Villar-Garea, M. Boix-Chornet, J. Espada, G. Schotta, T. Bonaldi, C. Haydon, S. Ropero, K. Petrie, N. G. Iyer, A. Pérez-Rosado, E. Calvo, J. A. Lopez, A. Cano, M. J. Calasanz, D. Colomer, M. Á. Piris, N. Ahn, A. Imhof, C. Caldas, T. Jenuwein, and M. Esteller. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature Genetics*, 37(4):391–400, 2005.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. In *Proceedings of the fourth annual international conference on Computational molecular biology - RECOMB '00*, pages 127–135, 2000.

R. Frisch and F. V. Waugh. Partial Time Regressions as Compared with Individual Trends. *Econometrica*, 1(4):387, 1933.

H. Fröhlich. biRte: Bayesian inference of context-specific regulator activities and transcriptional networks. *Bioinformatics*, 31(20):3290–3298, 2015.

C. W. Fuller, L. R. Middendorf, S. A. Benner, G. M. Church, T. Harris, X. Huang, S. B. Jovanovich, J. R. Nelson, J. A. Schloss, D. C. Schwartz, and D. V. Vezenov. The challenges of sequencing by synthesis. *Nature Biotechnology*, 27(11):1013–1023, 2009.

T. S. Furey. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, 13(12):840–852, 2012.

P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature reviews. Cancer*, 4(3): 177–183, 2004.

S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón, A. Medina-Rivera, H. Solano-Lira, C. Bonavides-Martínez, E. Pérez-Rueda, S. Alquicira-Hernández, L. Porrón-Sotelo, A. López-Fuentes, A. Hernández-Koutoucheva, V. Del Moral-Chavez, F. Rinaldi, and J. Collado-Vides. RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44:D133–D143, 2016.

A. Ganju, S. Khan, B. B. Hafeez, S. W. Behrman, M. M. Yallapu, S. C. Chauhan, and M. Jaggi. miRNA nanotherapeutics for cancer. *Drug Discovery Today*, 22(2): 424–432, 2017.

F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC bioinformatics*, 5:31, 2004.

L. Garcia-Alonso, C. H. Holland, M. M. Ibrahim, D. Turei, and J. Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29(8):1363–1375, 2019.

G. Garcia-Manero. Demethylating agents in myeloid malignancies. *Current Opinion in Oncology*, 20(6):705–710, 2008.

J. Gauthier, A. T. Vincent, S. J. Charette, and N. Derome. A brief history of bioinformatics. *Briefings in Bioinformatics*, pages 1–16, aug 2018. URL `https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby063/5066445`.

G. Geeven, R. E. van Kesteren, A. B. Smit, and M. C. de Gunst. Identification of context-specific gene regulatory networks with GEMULA-gene expression modeling using LAsso. *Bioinformatics*, 28(2):214–221, 2012.

E. I. George and R. E. Mcculloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and M. Snyder. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012.

J. Gillis and P. Pavlidis. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3):e1002444, 2012.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10(JUN), 2019.

X. Gong, P. Jia, and Z. Zhao. Investigating microRNA-transcription factor mediated regulatory network in glioblastoma. *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pages 258–263, 2010.

S. Goodwin, J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M. Schatz, and W. R. McCombie. Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome. *Genome Research*, 2015.

S. Gopalakrishnan, B. O. Van Emburgh, and K. D. Robertson. DNA methylation in development and human disease. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 647(1-2):30–38, 2008.

J. B. Greer and D. C. Whitcomb. Role of BRCA1 and BRCA2 mutations in pancreatic cancer. *Gut*, 56(5):601–605, 2007.

S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34:D140–144, 2006.

H. Gronemeyer, J. Å. Gustafsson, and V. Laudet. Principles for modulation of the nuclear receptor superfamily. *Nature Reviews Drug Discovery*, 3(11):950–964, 2004.

J. L. Gross and J. Yellen. *Handbook of Graph Theory*. CRC Press, 2003.

C. Grunau. MethDB–a public database for DNA methylation data. *Nucleic Acids Research*, 29(1):270–274, 2001.

H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, 2010.

B. Gurel, T. Iwata, C. M. Koh, S. Yegnasubramanian, W. G. Nelson, and A. M. De Marzo. Molecular alterations in prostate cancer as diagnostic, prognostic, and therapeutic targets. *Advances in Anatomic Pathology*, 15(6):319–331, 2008.

F. Gwinner, G. Boulday, C. Vandiedonck, M. Arnould, C. Cardoso, I. Nikolayeva, O. Guitart-Pla, C. V. Denis, O. D. Christophe, J. Beghain, E. Tournier-Lasserve, and B. Schwikowski. Network-based analysis of omics data: The LEAN method. *Bioinformatics*, 33(5):701–709, 2017.

M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33 (14):i37–i48, 2017.

G. T. Hart, A. K. Ramani, and E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120, 2006.

A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 422–433, 2001.

Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, 2017.

K. Hatano, B. Kumar, Y. Zhang, J. B. Coulter, M. Hedayati, B. Mears, X. Ni, T. A. Kudrolli, W. H. Chowdhury, R. Rodriguez, T. L. DeWeese, and S. E. Lupold. A functional screen identifies miRNAs that inhibit DNA repair and sensitize prostate cancer cells to ionizing radiation. *Nucleic Acids Research*, 43(8):4075–4086, 2015.

D. Hebenstreit. Methods, challenges and potentials of single cell RNA-seq. *Biology*, 1 (3):658–667, 2012.

M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. Gene regulatory network inference: Data integration in dynamic models-A review. *Biosystems*, 96(1): 86–103, 2009.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3):197–243, 1995.

D. Hernández-Lobato, J. M. Hernández-Lobato, and A. Suárez. Expectation Propagation for microarray data classification. *Pattern Recognition Letters*, 31(12): 1618–1626, 2010.

K. S. Hoek, N. C. Schlegel, P. Brafford, A. Sucker, S. Ugurel, R. Kumar, B. L. Weber, K. L. Nathanson, D. J. Phillips, M. Herlyn, D. Schadendorf, and R. Dummer. Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. *Pigment Cell Research*, 19(4):290–302, 2006.

P. Hogeweg. The roots of bioinformatics in theoretical biology. *PLoS Computational Biology*, 7(3), 2011.

A. Höglund, L. M. Nilsson, S. V. Muralidharan, L. A. Hasvold, P. Merta, M. Rudelius, V. Nikolova, U. Keller, and J. A. Nilsson. Therapeutic implications for the induced levels of Chk1 in Myc-expressing cancer cells. *Clinical Cancer Research*, 17: 7067–7079, 2011.

L. S. Hsu, H. C. Lee, G. Y. Chau, P. H. Yin, C. W. Chi, and W. Y. Lui. Aberrant methylation of EDNRB and p16 genes in hepatocellular carcinoma (HCC) in Taiwan. *Oncology Reports*, 15(2):507–511, 2006.

S. D. Hsu, F. M. Lin, W. Y. Wu, C. Liang, W. C. Huang, W. L. Chan, W. T. Tsai, G. Z. Chen, C. J. Lee, C. M. Chiu, C. H. Chien, M. C. Wu, C. Y. Huang, A. P. Tsou, and H. D. Huang. MiRTarBase: A database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 39:D163–169, 2011.

J. C. Huang, T. Babak, T. W. Corson, G. Chua, S. Khan, B. L. Gallie, T. R. Hughes, B. J. Blencowe, B. J. Frey, and Q. D. Morris. Using expression profiling data to identify human microRNA targets. *Nature methods*, 4(12):1045–1049, 2007.

S. Huang, K. Chaudhary, and L. X. Garmire. More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 8:84, 2017.

E. Hubbell, W. M. Liu, and R. Mei. Robust estimators for expression analysis. *Bioinformatics*, 18:1585–1592, 2002.

V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776, 2010.

A. Isomura and R. Kageyama. Ultradian oscillations and pulses: coordinating cellular responses and cell fate decisions. *Development*, 141(19):3627–3636, oct 2014.

R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(3S):245–254, 2003.

M. Jargosch, S. Kröger, E. Gralinska, U. Klotz, Z. Fang, W. Chen, U. Leser, J. Selbig, D. Groth, and R. Baumgrass. Data integration for identification of important transcription factors of STAT6-mediated cell fate decisions. *Genetics and Molecular Research*, 15(2), 2016.

N. Jayaram, D. Usvyat, and A. C. R. Martin. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, 2016.

D. W. Je, Y. M. O, Y. G. Ji, Y. Cho, and D. H. Lee. The inhibition of SRC family kinase suppresses pancreatic cancer cell proliferation, migration, and invasion. *Pancreas*, 43(5):768–76, 2014.

P. Jiang, M. L. Freedman, J. S. Liu, and X. S. Liu. Inference of transcriptional regulation in cancers. *Proceedings of the National Academy of Sciences*, 112(25): 7731–7736, 2015.

Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu. miR2Disease: A manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research*, 37:98–104, 2009.

L. Jin, X. Y. Zuo, W. Y. Su, X. L. Zhao, M. Q. Yuan, L. Z. Han, X. Zhao, Y. D. Chen, and S. Q. Rao. Pathway-based analysis tools for complex diseases: A Review. *Genomics, Proteomics and Bioinformatics*, 12(5):210–220, 2014.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007.

A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

H. B. Kang, H. R. Lee, D. J. Jee, S. H. Shin, S. S. Nah, S. Y. Yoon, and J. W. Kim. PRDM1, a Tumor-Suppressor Gene, is Induced by Genkwadaphnin in Human Colon Cancer SW620 Cells. *Journal of Cellular Biochemistry*, 117(1):172–179, 2016.

M. Kasowski, S. Kyriazopoulou-Panagiotopoulou, F. Grubert, J. B. Zaugg, A. Kundaje, Y. Liu, A. P. Boyle, Q. C. Zhang, F. Zakharia, D. V. Spacek, J. Li, D. Xie, A. Olarerin-George, L. M. Steinmetz, J. B. Hogenesch, M. Kellis, S. Batzoglou, and M. Snyder. Extensive variation in chromatin states across humans. *Science*, 342(6159):750–752, 2013.

S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–467, 1969.

A. Kel, U. Boyarskikh, P. Stegmaier, L. S. Leskov, A. V. Sokolov, I. Yevshin, N. Mandrik, D. Stelmashenko, J. Koschmann, O. Kel-Margoulis, M. Krull, A. Martínez-Cardús, S. Moran, M. Esteller, F. Kolpakov, M. Filipenko, and E. Wingender. Walking pathways with positive feedback loops reveal DNA methylation biomarkers of colorectal cancer. *BMC Bioinformatics*, 20(4), 2019.

I. M. Keseler, A. Mackie, A. Santos-Zavaleta, R. Billington, C. Bonavides-Martínez, R. Caspi, C. Fulcher, S. Gama-Castro, A. Kothari, M. Krummenacker, M. Latendresse, L. Muñiz-Rascado, Q. Ong, S. Paley, M. Peralta-Gil, P. Subhraveti, D. A. Velázquez-Ramírez, D. Weaver, J. Collado-Vides, I. Paulsen, and P. D. Karp. The EcoCyc database: Reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Research*, 41, 2017.

F. Khan. Genetic disorders and gene therapy. In *Biotechnology in Medical Sciences*, pages 237–262. CRC Press, 2014.

J. K. Kim, M. Samaranayake, and S. Pradhan. Epigenetic mechanisms in mammals. *Cellular and Molecular Life Sciences*, 66(4):596, 2008.

K. Kim, K. Jiang, S. L. Teng, L. J. Feldman, and H. Huang. Using biologically interrelated experiments to identify pathway genes in arabidopsis. *Bioinformatics*, 28 (6):815–822, 2012.

U. Klein, Y. Tu, G. A. Stolovitzky, M. Mattioli, G. Cattoretti, H. Husson, A. Freedman, G. Inghirami, L. Cro, L. Baldini, A. Neri, A. Califano, and R. Dalla-Favera. Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *The Journal of experimental medicine*, 194(11):1625–1638, 2001.

D. A. Kleinjan and V. van Heyningen. Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease. *The American Journal of Human Genetics*, 76(1):8–32, 2005.

B. Klinger and N. Blüthgen. Reverse engineering gene regulatory networks by modular response analysis – a benchmark. *Essays In Biochemistry*, 62(4):535–547, 2018.

S. Komaki, Y. Shiwa, R. Furukawa, T. Hachiya, H. Ohmomo, R. Otomo, M. Satoh, J. Hitomi, K. Sobue, M. Sasaki, and A. Shimizu. iMETHYL: an integrative database

of human DNA methylation, gene expression, and genomic variation. *Human Genome Variation*, 5(1), 2018.

S. Komili and P. A. Silver. Coupling and coordination in gene expression processes: A systems biology view. *Nature Reviews Genetics*, 9:38–48, 2008.

S. Konishi, T. Ando, and S. Imoto. Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91(1):27–43, 2004.

M. Kozak. Regulation of translation via mrna structure in prokaryotes and eukaryotes. *Gene*, 361:13–37, 2005.

A. Krämer, J. Green, J. Pollard, and S. Tugendreich. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics (Oxford, England)*, 30(4):523–530, 2014.

K. Krishnan, A. L. Steptoe, H. C. Martin, S. Wani, K. Nones, N. Waddell, M. Mariasegaram, P. T. Simpson, S. R. Lakhani, B. Gabrielli, A. Vlassov, N. Cloonan, and S. M. Grimmond. MicroRNA-182-5p targets a network of genes involved in DNA repair. *Rna*, 19(2):230–242, 2013.

A. Lachmann, H. Xu, J. Krishnan, S. I. Berger, A. R. Mazloom, and A. Ma'ayan. ChEA: Transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, 26(19):2438–2444, 2010.

M. Lambert, S. Jambon, S. Depauw, and M. H. David-Cordonnier. Targeting transcription factors for cancer treatment. *Molecules*, 23(6), 2018a.

S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. The Human Transcription Factors. *Cell*, 172(4):650–665, 2018b.

T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. 'T Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. Van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. Macarthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, Á. Carracedo, S. E. Antonarakis, R. Häsler, A. C. Syvänen, G. J. Van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

M. Lauriola, G. Ugolini, S. Rivetti, S. Nanì, G. Rosati, S. Zanotti, I. Montroni, A. Manaresi, D. Zattoni, A. Belluzzi, L. Castellani, G. D'Uva, G. Mattei, M. Taffurelli, P. Strippoli, and R. Solmi. IL23R, NOD2/CARD15, ATG16L1 and PHOX2B polymorphisms in a group of patients with Crohn's disease and correlation

with sub-phenotypes. *International Journal of Molecular Medicine*, 27(3):469–477, 2011.

M. Leddin, C. Perrod, M. Hoogenkamp, S. Ghani, S. Assi, S. Heinz, N. K. Wilson, G. Follows, J. Schönheit, L. Vockentanz, A. M. Mosammam, W. Chen, D. G. Tenen, D. R. Westhead, B. Göttgens, C. Bonifer, and F. Rosenbauer. Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells. *Blood*, 117(10):2827–2838, 2011.

H. J. Lee, A. W. Wark, and R. M. Corn. Microarray methods for protein biomarker detection. *Analyst*, 133(8):975–983, 2008.

C. Lefebvre, P. Rajbhandari, M. J. Alvarez, P. Bandaru, W. K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B. C. Bisikirska, K. Basso, P. Beltrao, N. Krogan, J. Gautier, R. Dalla-Favera, and A. Califano. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, 6, 2010.

R. Lehmann, L. Childs, P. Thomas, M. Abreu, L. Fuhr, H. Herzel, U. Leser, and A. Relógio. Assembly of a comprehensive regulatory network for the mammalian circadian clock: A bioinformatics approach. *PLoS ONE*, 10(5), 2015.

B. Lemon and R. Tjian. Orchestrated response: A symphony of transcription factors for gene control. *Genes and Development*, 14(20):2551–2569, 2000.

T. L. Lenstra and F. C. Holstege. The discrepancy between chromatin factor location and effect. *Nucleus (United States)*, 3(3), 2012.

H. Li, Y. Du, D. Zhang, L. N. Wang, C. Yang, B. Liu, W. J. Wang, L. Shi, W. G. Hong, L. Zhang, and Y. X. Yang. Identification of novel DNA methylation markers in colorectal cancer using MIRA-based microarrays. *Oncology Reports*, 28(1):99–104, 2012.

K. C. Li, A. Palotie, S. Yuan, D. Bronnikov, D. Chen, X. Wei, O. W. Choi, J. Saarela, and L. Peltonen. Finding disease candidate genes by liquid association. *Genome Biology*, 8(10):R205, 2007a.

L. Li and J. R. Davie. The role of Sp1 and Sp3 in normal and cancer cell biology. *Annals of Anatomy*, 192(5):275–283, 2010.

P. Li, C. Zhang, E. J. Perkins, P. Gong, and Y. Deng. Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. *BMC bioinformatics*, 8 Suppl 7:S13, 2007b.

R. Li, F. Liang, M. Li, D. Zou, S. Sun, Y. Zhao, W. Zhao, Y. Bao, J. Xiao, and Z. Zhang. MethBank 3.0: A database of DNA methylomes across a variety of species. *Nucleic Acids Research*, 46(D1):D288–D295, 2018.

Y. Li, M. Liang, and Z. Zhang. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS computational biology*, 10, 2014.

L. Liang, L. Gao, X. P. Zou, M. L. Huang, G. Chen, J. J. Li, and X. Y. Cai. Diagnostic significance and potential function of miR-338-5p in hepatocellular carcinoma: A bioinformatics study with microarray and RNA sequencing data. *Molecular Medicine Reports*, 17(2):2297–2312, 2018.

S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, pages 18–29, 1998.

Y. Lichtblau, K. Zimmermann, B. Haldemann, D. Lenze, M. Hummel, and U. Leser. Comparative assessment of differential network analysis methods. *Briefings in bioinformatics*, 18(5):837–850, 2017.

C. Liu, C. E. Banister, C. C. Weige, D. Altomare, J. H. Richardson, C. M. Contreras, and P. J. Buckhaults. PRDM1 silences stem cell-related genes and inhibits proliferation of human colon tumor organoids. *Proceedings of the National Academy of Sciences of the United States of America*, 115(22):E5066–E5075, 2018.

H. Liu, Z. Pan, A. Li, S. Fu, Y. Lei, H. Sun, M. Wu, and W. Zhou. Roles of chemokine receptor 4 (CXCR4) and chemokine ligand 12 (CXCL12) in metastasis of hepatocellular carcinoma cells. *Cellular & molecular immunology*, 5(5):373–378, 2008.

H. Liu, P. D'Andrade, S. Fulmer-Smentek, P. Lorenzi, K. W. Kohn, J. N. Weinstein, Y. Pommier, and W. C. Reinhold. mRNA and microRNA Expression Profiles of the NCI-60 Integrated with Drug Activities. *Mol Cancer Ther*, 9(5):1080–1091, 2010.

J. Liu, Y. Chi, C. Zhu, and Y. Jin. A time series driven decomposed evolutionary optimization approach for reconstructing large-scale gene regulatory networks based on fuzzy cognitive maps. *BMC Bioinformatics*, 18(1), 2017.

Z. P. Liu, C. Wu, H. Miao, and H. Wu. RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, 2015:1–12, 2015.

S. Lou, H.-M. Lee, H. Qin, J.-W. Li, Z. Gao, X. Liu, L. L. Chan, V. Lam, W.-Y. So, Y. Wang, S. Lok, J. Wang, R. C. Ma, S. K. Tsui, J. Chan, T.-F. Chan, and K. Y. Yip. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome biology*, 15(7):408, 2014.

P. J. F. Lucas. Restricted Bayesian Network Structure Learning. In J. Gâmez, S. Moral, and A. Salmeron, editors, *Advances in Bayesian Networks*, pages 217–234. Springer, Berlin, 2004.

X. Luo and Y. Wei. Nonparametric Bayesian learning of heterogeneous dynamic transcription factor networks. *Annals of Applied Statistics*, 12(3):1749–1772, 2018.

N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, 2004.

E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5): 1202–1214, 2015.

F. Magdinier, S. Ribieras, G. M. Lenoir, L. Frappart, and R. Dante. Down-regulation of BRCA1 in human sporadic breast cancer; analysis of DNA methylation patterns of the putative promoter region. *Oncogene*, 17(24):3169–3176, 1998.

S. Mamlouk, L. H. Childs, D. Aust, D. Heim, F. Melching, C. Oliveira, T. Wolf, P. Durek, D. Schumacher, H. Bläker, M. Von Winterfeld, B. Gastl, K. Möhr, A. Menne, S. Zeugner, T. Redmer, D. Lenze, S. Tierling, M. Möbs, W. Weichert, G. Folprecht, E. Blanc, D. Beule, R. Schäfer, M. Morkel, F. Klauschen, U. Leser, and C. Sers. DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer. *Nature Communications*, 8, 2017.

D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229–239, 2009.

D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, T. D. Consortium, M. Kellis, J. J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9:796–804, 2012.

A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7 Suppl 1:S7, 2006.

F. Markowetz and R. Spang. Inferring cellular networks-a review. *BMC bioinformatics*, 8 Suppl 6:S5, 2007.

F. Markowetz, D. Kostka, O. G. Troyanskaya, and R. Spang. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23(13):i305–i312, 2007.

L. Martignetti, L. Calzone, E. Bonnet, E. Barillot, and A. Zinovyev. ROMA: Representation and quantification of module activity from target expression data. *Frontiers in Genetics*, 7:18, 2016.

M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin,

S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337:1190–1195, 2012.

M. W. Mayo and A. S. Baldwin. The transcription factor NF-kappaB: control of oncogenesis and cancer therapy resistance. *Biochimica et biophysica acta*, 1470(2): M55–M62, 2000.

P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19 suppl 2:ii122–ii129, 2003.

V. Menon, S. Yarahmadian, and V. Rezania. Novel EM based ML Kalman estimation framework for superresolution of stochastic three-states microtubule signal. *BMC Systems Biology*, 12(6), 2018.

P. Meyer, T. Cokelaer, D. Chandran, K. H. Kim, P.-R. Loh, G. Tucker, M. Lipson, B. Berger, C. Kreutz, A. Raue, B. Steiert, J. Timmer, E. Bilal, Dream 6 7 Parameter Estimation Consortium, H. M. Sauro, G. Stolovitzky, and J. Saez-Rodriguez. Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. *BMC systems biology*, 8(1):13, jan 2014.

P. E. Meyer, F. Lafitte, and G. Bontempi. Minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9, 2008.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. MacAulay, W. Jawaid, E. Diamanti, S. I. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher, and B. Göttgens. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, 33(3):269–276, 2015.

M. Morkel, P. Riemer, H. Bläker, and C. Sers. Similar but different: Distinct roles for KRAS and BRAF oncogenes in colorectal cancer development and therapy resistance. *Oncotarget*, 6(25):20785–20800, 2015.

B. Munsky, G. Neuert, and A. Van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

J. Musa, M.-M. Aynaud, O. Mirabeau, O. Delattre, and T. G. Grünewald. MYBL2 (B-Myb): a central regulator of cell proliferation, cell survival and differentiation involved in tumorigenesis. *Cell Death and Disease*, 8:e2895, 2017.

U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 2008.

J. R. Naranjo, H. Zhang, D. Villar, P. González, X. M. Dopazo, J. Morón-Oset, E. Higueras, J. C. Oliveros, M. D. Arrabal, A. Prieto, P. Cercós, T. González, A. De La Cruz, J. Casado-Vela, A. Rábano, C. Valenzuela, M. Gutierrez-Rodriguez, J. Y. Li, and B. Mellström. Activating transcription factor 6 derepression mediates neuroprotection in Huntington disease. *Journal of Clinical Investigation*, 126(2): 627–638, 2016.

National Cancer Institute Wiki. Cancer Gene Index End User Documentation, 2014. URL `https://wiki.nci.nih.gov/x/hC5yAQ`. Accessed 14.07.2016.

V. A. Naumov, E. V. Generozov, N. B. Zaharjevskaya, D. S. Matushkina, A. K. Larin, S. V. Chernyshov, M. V. Alekseev, Y. A. Shelygin, and V. M. Govorun. Genome-scale analysis of DNA methylation in colorectal cancer using Infinium HumanMethylation450 BeadChips. *Epigenetics*, 8(9):921–934, 2013.

A. Nekrutenko and J. Taylor. Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 9(13): 667–672, 2012.

N. C. Nicolaides, I. Correa, C. Casadevall, S. Travali, K. J. Soprano, and B. Calabretta. The Jun family members, c-Jun and JunD, transactivate the human c-myb promoter via an Ap1-like element. *Journal of Biological Chemistry*, 267:19665–19672, 1992.

A. Nishiyama, L. Xin, A. A. Sharov, M. Thomas, G. Mowrer, E. Meyers, Y. Piao, S. Mehta, S. Yee, Y. Nakatake, C. Stagg, L. Sharova, L. S. Correa-Cerro, U. Bassey, H. Hoang, E. Kim, R. Tapnio, Y. Qian, D. Dudekula, M. Zalzman, M. Li, G. Falco, H. T. Yang, S. L. Lee, M. Monti, I. Stanghellini, M. N. Islam, R. Nagaraja, I. Goldberg, W. Wang, D. L. Longo, D. Schlessinger, and M. S. Ko. Uncovering Early Response of Gene Regulatory Networks in ESCs by Systematic Induction of Transcription Factors. *Cell Stem Cell*, 5:420–433, 2009.

A. Noor, A. Ahmad, and E. Serpedin. SparseNCA: Sparse Network Component Analysis for Recovering Transcription Factor Activities with Incomplete Prior Information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(2):387–395, 2018.

J. Nourse, J. Braun, K. Lackner, S. Hüttelmaier, and S. Danckwardt. Large-scale identification of functional microRNA targeting reveals cooperative regulation of the hemostatic system. *Journal of Thrombosis and Haemostasis*, 16(11):2233–2245, 2018.

G. Nuel. Apprentissage de Motif. Seminar 4 of lecture Modèles aléatoires en vue de la biologie, Université Paris Descartes, 2013. Lecture notes.

M. Ongenaert, L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert, and W. Van Criekinge. PubMeth: A cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Research*, 36(Database Issue):D842–D846, 2008.

R. Opgen-Rhein and K. Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology*, 1:37, 2007.

G. Pan, J. Li, Y. Zhou, H. Zheng, and D. Pei. A negative feedback loop of transcription factors that controls stem cell pluripotency and self-renewal. *The FASEB Journal*, 20(10):1730–1732, 2006.

N. Parikh, S. Hilsenbeck, C. J. Creighton, T. Dayaram, R. Shuck, E. Shinbrot, L. Xi, R. A. Gibbs, D. A. Wheeler, and L. A. Donehower. Effects of TP53 mutational status on gene expression patterns across 10 human cancer types. *Journal of Pathology*, 232(5):522–533, 2014.

J. D. Partridge, C. Scott, Y. Tang, R. K. Poole, and J. Green. Escherichia coli transcriptome dynamics during the transition from anaerobic to aerobic conditions. *Journal of Biological Chemistry*, 281:27806–27815, 2006.

A. Pataskar and V. K. Tiwari. Computational challenges in modeling gene regulatory events. *Transcription*, 7(5):188–195, 2016.

A. Petitjean, M. I. Achatz, A. L. Borresen-Dale, P. Hainaut, and M. Olivier. TP53 mutations in human cancers: Functional selection and impact on cancer prognosis and outcomes. *Oncogene*, 26(15):2157–2165, 2007.

T. Picchetti, J. Chiquet, M. Elati, P. Neuvial, R. Nicolle, and E. Birmelé. A model for gene deregulation detection using expression data. *BMC Systems Biology*, 9(6), 2015.

J. K. Pickrell, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*, 27(15):2144–2146, 2011.

A. Pombo and N. Dillon. Three-dimensional genome architecture: Players and mechanisms. *Nature Reviews Molecular Cell Biology*, 16(4):245–257, 2015.

N. Rajewsky. microRNA target predictions in animals. *Nature genetics*, 38 Suppl: S8–13, 2006.

S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–80, 2014.

D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabad, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A.

Smith, C. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.

J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Molecular cell*, 2015.

H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, 2007.

G. Romano and L. N. Kwong. Diagnostic and therapeutic applications of miRNA-based strategies to cancer immunotherapy. *Cancer and Metastasis Reviews*, 37(1):45–53, 2018.

D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics*, 24 (3):227–235, 2000.

C. Rubie, V. O. Frick, M. Wagner, C. Weber, B. Kruse, K. Kempf, J. König, B. Rau, and M. Schilling. Chemokine expression in hepatocellular carcinoma versus colorectal liver metastases. *World journal of gastroenterology : WJG*, 12(41):6627–33, 2006.

J. Rung and A. Brazma. Reuse of public genome-wide gene expression data. *Nat Rev Genet*, 14:89–99, 2013.

T. Rydén. Em versus Markov chain monte carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis*, 3(4):659–688, 2008.

S. Sadasivam, S. Duan, and J. A. DeCaprio. The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression. *Genes and Development*, 26: 474–489, 2012.

M. Sadelain, E. P. Papapetrou, and F. D. Bushman. Safe harbours for the integration of new DNA in the human genome. *Nat Rev Cancer*, 12(1):51–58, 2012.

A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32:D91–94, 2004.

N. V. Sankpal, T. P. Fleming, P. K. Sharma, H. J. Wiedner, and W. E. Gillanders. A double-negative feedback loop between EpCAM and ERK contributes to the regulation of epithelial-mesenchymal transition in cancer. *Oncogene*, 36:3706–3717, 2017.

H. M. Sauro. Control and regulation of pathways via negative feedback. *Journal of the Royal Society Interface*, 14(127), 2017.

T. Schacht, M. Oswald, R. Eils, S. B. Eichmüller, and R. König. Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics*, 30(17):i401–i407, 2014.

T. Schaffter, D. Marbach, and D. Floreano. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 1995.

S. Schmeier, T. Alam, M. Essack, and V. B. Bajic. TcoF-DB v2: Update of the database of human and mouse transcription co-factors and transcription factor interactions. *Nucleic Acids Research*, 45:D145–D150, 2017.

M. H. Schulz, W. E. Devanny, A. Gitter, S. Zhong, J. Ernst, and Z. Bar-Joseph. DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Systems Biology*, 6, 2012.

E. A. Semenova, M. chul Kwon, K. Monkhorst, J. Y. Song, R. Bhaskaran, O. Krijgsman, T. Kuilman, D. Peters, W. A. Buikhuisen, E. F. Smit, C. Pritchard, M. Cozijnsen, J. van der Vliet, J. Zevenhoven, J. P. Lambooij, N. Proost, E. van Montfort, A. Velds, I. J. Huijbers, and A. Berns. Transcription Factor NFIB Is a Driver of Small Cell Lung Cancer Progression in Mice and Marks Metastatic Disease in Patients. *Cell Reports*, 16(3):631–643, 2016.

Q. Shi, C. Zhang, W. Guo, T. Zeng, L. Lu, Z. Jiang, Z. Wang, J. Liu, and L. Chen. Local network component analysis for quantifying transcription factor activities. *Methods*, 124:25–35, 2017.

R. H. Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nature reviews. Cancer*, 6(10):813–823, 2006.

S. Sikdar and S. Datta. A novel statistical approach for identification of the master regulator transcription factor. *BMC Bioinformatics*, 18(1):79, 2017.

J. C. Spall. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, 2005. doi: 10.1002/0471722138.

F. Spitz and E. E. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012.

E. Steele, A. Tucker, P. A. 't Hoen, and M. J. Schuemie. Literature-based priors for gene regulatory networks. *Bioinformatics*, 25(14):1768–1774, 2009.

G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. Iny Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet. The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, pages 1.30.1–1.30.33, 2016.

S. H. Sternberg and J. A. Doudna. Expanding the Biologist's Toolkit with CRISPR-Cas9. *Molecular Cell*, 58(4):568–574, 2015.

T. Stiewe, S. Tuve, M. Peter, A. Tannapfel, A. H. Elmaagacli, and B. M. Pützer. Quantitative TP73 Transcript Analysis in Hepatocellular Carcinomas. *Clinical Cancer Research*, 10(2):626–633, 2004.

K. P. Stim-Herndon, T. M. Flores, and G. N. Bennett. Molecular characterization of adiY, a regulatory gene which affects expression of the biodegradative acid-induced arginine decarboxylase gene (adiA) of Escherichia coli. *Microbiology*, 142:1311–1320, 1996.

J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.

A. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein encoding transcriptomes. *Proceedings of the National Academy of Sciences*, 101(16):6062–6067, 2004.

R. u. Takahashi, M. Prieto-Vila, I. Kohama, and T. Ochiya. Development of miRNA-based therapeutic approaches for cancer patients. *Cancer Science*, 110(4): 1140–1147, 2019.

H. Tan and X. Zhou. Detection of combinatorial mutational patterns in human cancer genomes by exclusivity analysis. *Methods in Molecular Biology*, 1711:3–11, 2018.

V. B. Teif and K. Rippe. Statistical-mechanical lattice models for protein-DNA binding in chromatin. *Journal of Physics Condensed Matter*, 22(41), 2010.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455 (7216):1061–1068, 2008.

The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine*, 368 (22):2059–2074, 2013.

P. Thomas, P. Durek, I. Solt, B. Klinger, F. Witzel, P. Schulthess, Y. Mayer, D. Tikk, N. Blüthgen, and U. Leser. Computer-assisted curation of a human regulatory core network from the biological literature. *Bioinformatics*, 31(8):1258–1266, 2015.

S. A. Thomas and Y. Jin. Reconstructing biological gene regulatory networks: Where optimization meets big data. *Evolutionary Intelligence*, 7(1):29–47, 2014.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics*, 1963.

M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. a. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23(1):137–144, 2005.

S. Trescher and U. Leser. Estimation of Transcription Factor Activity in Knockdown Studies. *Scientific Reports*, 9(1):9593, 2019.

S. Trescher, J. Münchmeyer, and U. Leser. Estimating genome-wide regulatory activity from multi-omics data sets using mathematical optimization. *BMC Systems Biology*, 11(1):1–18, 2017.

S. Valverde, S. Ohse, M. Turalska, B. J. West, and J. Garcia-Ojalvo. Structural determinants of criticality in biological networks. *Frontiers in Physiology*, 6:127, 2015.

N. L. van Berkum, E. Lieberman-Aiden, L. Williams, M. Imakaev, A. Gnirke, L. A. Mirny, J. Dekker, and E. S. Lander. Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. *Journal of Visualized Experiments*, 39:1869, 2010.

T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, K. Marchal, T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal. {SynTReN}: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1):43, 2006.

M. van Kouwenhove, M. Kedde, and R. Agami. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nature reviews. Cancer*, 11(9):644–656, 2011.

J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–263, 2009.

C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):237–245, 2010.

C. E. Vejnar and E. M. Zdobnov. MiRmap: Comprehensive prediction of microRNA target repression strength. *Nucleic Acids Research*, 40(22):11673–11683, 2012.

S. Vlaic, W. Schmidt-Heck, M. Matz-Soja, E. Marbach, J. Linde, A. Meyer-Baese, S. Zellmer, R. Guthke, and R. Gebhardt. The extended TILAR approach: A novel tool for dynamic modeling of the transcription factor network regulating the adaption to in vitro cultivation of murine hepatocytes. *BMC Systems Biology*, 6, 2012.

K. V. Voelkerding, S. A. Dames, and J. D. Durtschi. Next-generation sequencing:from basic research to diagnostics, 2009.

B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz Jr., and K. W. Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, 2013.

G. Vorbrueggen, F. Kalkbrenner, S. Guehmann, and K. Moelling. The carboxyterminus of human c-myb protein stimulates activated transcription in trans. *Nucleic Acids Research*, 22:2466–2475, 1994.

W. Wang, V. Baladandayuthapani, J. S. Morris, B. M. Broom, G. Manyam, and K. A. Do. IBAG: Integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159, 2013.

Y. X. Wang and H. Huang. Review on statistical methods for gene network reconstruction using expression data. *Journal of Theoretical Biology*, 362:53–61, 2014.

S. Washietl, J. S. Pedersen, J. O. Korbel, C. Stocsits, A. R. Gruber, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Reiche, A. Tanzer, C. Ucla, C. Wyss, S. E. Antonarakis, F. Denoeud, J. Lagarde, J. Drenkow, P. Kapranov, T. R. Gingeras, R. Guigó, M. Snyder, M. B. Gerstein, A. Reymond, I. L. Hofacker, and P. F. Stadler. Structured rnas in the encode selected regions of the human genome. *Genome Research*, 17(6):852–864, 2007.

J. Watson and F. Crick. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

F. Watt and P. L. Molloy. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes & development*, 2(9):1136–1143, 1988.

J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–20, 2013.

K. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), 2019. URL `www.genome.gov/sequencingcostsdata`. Accessed 02.09.2019.

M. Wilson, J. Brosens, H. Schwenen, and E. Lam. FOXO and FOXM1 in Cancer: The FOXO-FOXM1 Axis Shapes the Outcome of Cancer Chemotherapy. *Current Drug Targets*, 12:1256–1266, 2011.

E. Wingender, P. Dietze, H. Karas, and R. Knüppel. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24(1): 238–241, 1996.

C. F. J. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

C. T. Wu and J. R. Morris. Genes, genetics, and epigenetics: A correspondence. *Science*, 293(5532):1103–1105, 2001.

M.-Y. Wu, X.-F. Zhang, D.-Q. Dai, L. Ou-Yang, Y. Zhu, and H. Yan. Regularized logistic regression with network-based pairwise interaction for biomarker identification in breast cancer. *BMC Bioinformatics*, 17(1):108, 2016.

J. Xi, M. Wang, and A. Li. Discovering mutated driver genes through a robust and sparse co-regularized matrix factorization framework with prior information from mRNA expression patterns and interaction network. *BMC bioinformatics*, 19(1):214, 2018.

W. Yan, W. Xue, J. Chen, and G. Hu. Biological networks for cancer candidate biomarkers discovery. *Cancer Informatics*, 15:1–7, 2016.

K. Yandell. An Array of Options - A guide for how and when to transition from the microarray to RNA-seq, 2015. URL `https://www.the-scientist.com/lab-tools/an-array-of-options-35381`. Accessed 02.09.2019.

X. Yang, X. Zu, J. Tang, W. Xiong, Y. Zhang, F. Liu, and Y. Jiang. Zbtb7 suppresses the expression of CDK2 and E2F4 in liver cancer cells: Implications for the role of Zbtb7 in cell cycle regulation. *Molecular Medicine Reports*, 5(6):1475–1480, 2012.

Y. L. Yang, J. Suen, M. P. Brynildsen, S. J. Galbraith, and J. C. Liao. Inferring yeast cell cycle regulators and interactions using transcription factor activities. *BMC Genomics*, 6(1):90, 2005.

A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D. Sheppard, K. Taylor,

A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino, and P. Flicek. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, 2016.

Y. Yin, J. Zhong, S. W. Li, J. Z. Li, M. Zhou, Y. Chen, Y. Sang, and L. Liu. TRIM11, a direct target of miR-24-3p, promotes cell proliferation and inhibits apoptosis in colon cancer. *Oncotarget*, 7(52):86755–86765, 2016.

N. Yosef, A. K. Shalek, J. T. Gaublomme, H. Jin, Y. Lee, A. Awasthi, C. Wu, K. Karwacz, S. Xiao, M. Jorgolli, D. Gennert, R. Satija, A. Shakya, D. Y. Lu, J. J. Trombetta, M. R. Pillai, P. J. Ratcliffe, M. L. Coleman, M. Bix, D. Tantin, H. Park, V. K. Kuchroo, and A. Regev. Dynamic regulatory network controlling T H 17 cell differentiation. *Nature*, 496(7446):461–468, 2013.

B. Zacher, K. Abnaof, S. Gade, E. Younesi, A. Tresch, and H. Fröhlich. Joint bayesian inference of condition-specific miRNA and transcription factor activities from combined gene and microRNA expression data. *Bioinformatics*, 28(13):1714–1720, 2012.

J. Zhang and S. Zhang. Modular Organization of Gene Regulatory Networks. In *Encyclopedia of Systems Biology*, pages 1437–1441. Springer New York, New York, NY, 2013.

X. Zhang, Q. L. Lv, Y. T. Huang, L. H. Zhang, and H. H. Zhou. Akt/FoxM1 signaling pathway-mediated upregulation of MYBL2 promotes progression of human glioma. *Journal of Experimental and Clinical Cancer Research*, 36, 2017.

S. Zhao, W. P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, 9(1), 2014.

Z. Zhu, H. Wang, Y. Wei, F. Meng, Z. Liu, and Z. Zhang. Downregulation of PRDM1 promotes cellular invasion and lung cancer metastasis. *Tumor Biology*, 39(4), 2017.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(2):301–320, 2005.

List of Figures

List of Tables

List of Acronyms

3C	Chromosome conformation capture
AML	Acute myeloid leukemia
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
areA	Analytic rank-based enrichment analysis
BiRte	Bayesian inference of context-specific regulator activities and transcriptional networks
cDNA	Complementary DNA
ChEA	ChIP Enrichment Analysis
ChIP	Chromatin immunoprecipitation
CNV	Copy number variation
COAD	Colon adenocarcinoma
DNA	Deoxyribonucleic acid
DoRothEA	Discriminant Regulon Expression Analysis
DPI	Data processing inequality
EML	Extreme machine learning
ENCODE	Encyclopedia of DNA Elements
FBL	Feedback loop
Floræ	Feedback loops in regulatory activity estimation
FPKM	Fragment per kilobase of exon per million mapped reads
GEO	Gene Expression Omnibus
GGM	Gaussian graphical model
GNF	Genomics Institute of the Novartis Research Foundation
GNW	GeneNetWeaver
GRN	Gene regulatory network
HGNC	HUGO Gene Nomenclature Committee
ISMARA	Integrated System for Motif Activity Response Analysis
KD	Knockdown
KO	Knockout
LAR	Least Angle Regression

LASSO	Least absolute shrinkage and selection operator
LEAN	Local enrichment analysis
LIHC	Liver hepatocellular carcinoma
MAPK	Mitogen-activated protein kinase
MCMC	Markov Chain Monte Carlo
MI	Mutual information
miRNA	MicroRNA
mRNA	Messenger RNA
NCA	Network component analysis
NCI	National Cancer Institute
NEM	Nested Effects Model
NGS	Next-generation sequencing
ODE	Ordinary differential equation
PAAD	Pancreatic adenocarcinoma
RABIT	Regression Analysis with Background Integration
RACER	Regression Analysis of Combined Expression Regulation
RBP	RNA-binding protein
RNA	Ribonucleic acid
ROMA	Representation and quantification of Module Activities
TAD	topologically associating domain
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TFBS	Transcription factor binding sites
tRNA	Transfer RNA
UCSC	University of California, Santa Cruz
UTR	Untranslated region
VIPER	Virtual inference of protein activity by enriched regulon analysis

Selbständigkeitserklärung

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß § 7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 42/2018 am 11.07.2018 angegebenen Hilfsmittel angefertigt habe.

Berlin, den 29.10.2019 Saskia Trescher